MASTER EN SCIENCES MATHÉMATIQUES

Analyse en composantes principales et analyse factorielle discriminante symboliques

Martin, Amélie

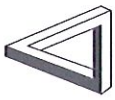
Award date:
2006

FUNDP
NAMUR

Faculté des Sciences
Département de Mathématique

Rempart de la Vierge, 8
B - 5000 Namur (Belgique)

Analyse en composantes principales et analyse factorielle discriminante symboliques



Mémoire présenté pour l'obtention
du grade de
Licencié en Sciences Mathématiques
par

MARTIN **Amélie**

Promoteur : HARDY André

Année Académique 2005-2006

Résumé

Dans de nombreux problèmes statistiques, nous devons travailler avec un nombre important de variables. Les méthodes factorielles nous permettent de diminuer ce nombre de variables afin d'obtenir une représentation graphique des données. Dans ce mémoire, nous étudierons deux de ces méthodes : l'analyse en composantes principales et l'analyse factorielle discriminante. Nous les développerons dans le cas classique et dans le cas symbolique et nous les appliquerons à l'aide du logiciel d'analyse de données symboliques Sodas.

ABSTRACT

In many statistical problems, we have to work with a considerable number of variables. Factorial methods allow us to reduce this number of variables in order to get a graphical representation of the data. In this thesis, we will study two methods : the principal component analysis and the factorial discriminant analysis. We will develop them in the classical case and in the symbolic case and we will apply them to the analysis of symbolic data with the Sodas software.

Remerciements

Je tiens tout d'abord à remercier Monsieur André Hardy pour m'avoir donné la possibilité de réaliser mon mémoire sous sa supervision et de m'avoir suivie tout au long de sa rédaction en me donnant de nombreux conseils qui m'ont permis de constamment améliorer ce travail.

Je remercie également Nathanael Kasoro pour ses nombreuses explications en ce qui concerne la création d'une base de données symboliques pour le logiciel Sodas, me permettant ainsi de réaliser de meilleures applications.

Je tiens tout particulièrement à remercier mes parents pour leur aide, leurs encouragements et leur soutien permanent tout au long de ces quatre années d'études. Merci également à mes grands parents.

Je remercie également ma tante Patricia Martin pour sa patience lors de la lecture de mon mémoire et les corrections qu'elle y a apporté.

Je tiens à adresser un merci tout spécial à Alain Mordant pour ses nombreuses aides au niveau informatique lors de la rédaction de ce mémoire et également pour sa relecture.

Je tiens enfin à remercier Vincent, Jérôme, Sébastien et Adeline pour leur soutien et leurs encouragements.

Table des matières

Introduction

Les statistiques cherchent à étudier les faits de la vie courante (qu'ils soient économiques, sociaux, politique, ...) à travers un ensemble de données reflétant au mieux ces différents faits. Pour cela, nous considérons des individus sur lesquels nous mesurons des variables, ce qui nous permet d'obtenir un ensemble de données sur lequel nous pouvons travailler.

Dans le cas le plus simple, lorsque nous mesurons une variable sur un individu, nous obtenons une et une seule valeur. Par exemple, le nombre d'enfants dans une famille sera un et un seul nombre. Nous appellerons alors les données que nous obtenons des *données classiques.*

Mais il arrive que, pour mesurer certains faits, une seule valeur ne suffise pas. Par exemple, si la variable correspond à la description de la couleur de certaines fleurs, une valeur ne suffit pas toujours. La valeur de la variable sur un individu est un ensemble de valeurs. Ces données seront appelées *données symboliques.*

Lorsque l'on étudie un ensemble de données, une des premières choses que nous souhaitons est d'avoir une représentation de ces données. Pour cela, nous réalisons un graphique dont les axes sont les différentes variables. Cependant, cela n'est possible que dans un espace de dimension 2 ou 3 ($\mathbb{R}^2$ ou $\mathbb{R}^3$) alors que, dans un grand nombre de problèmes statistiques que nous rencontrons, le nombre de variables mesurées est supérieur à 3, ce qui nous empêche de réaliser cette représentation.

Pour résoudre ce problème, des méthodes ont été mises au point afin de diminuer le nombre de variables intervenant dans la description des données, comme par exemple l'analyse en composantes principales (ACP) et l'analyse factorielle discriminante (AFD). Ces méthodes nous permettennt de ramener ce nombre à 2 ou 3 *sans perdre trop d'informations* sur les données et nous pouvons ainsi obtenir notre représentation.

L'analyse en composantes principales cherche à diminuer le nombre de variables d'un problème en tenant compte de la variabilité des données. Pour cela, nous chercherons un nombre inférieur de nouvelles variables (généralement 2 ou 3), que nous appellerons composantes principales, afin de représenter les données de départ dans l'espace formé par ces composantes principales.

Il arrive parfois que les données sur lesquelles nous travaillons forment des classes. Dans ce cas, nous utiliserons l'analyse factorielle discriminante qui cherche à diminuer le nombre de variables et qui tient en plus compte des classes que les différents objets peuvent former. Nous chercherons donc également de nouveaux axes que nous appellerons ici axes factoriels, qui nous permettront de retrouver ces classes dans l'espace de dimension inférieure.

Par exemple, si nous cherchons à étudier ce qui influence le prix d'une maison, nous allons définir un grand nombre de variables : la superficie totale de la maison, le nombre de pièces, les travaux à effectuer, la proximité de la ville, ... Nous ne saurons pas représenter directement les différentes maisons sur lesquelles nous avons mesuré ces variables sur un graphe. Nous devrons donc chercher de nouveaux axes afin de trouver de nouvelles coordonnées pour ces maisons. Nous utiliserons donc l'analyse en composantes principales.

Nous pourrions également vouloir tenir compte des quartiers où se situent ces maisons. Nous aurions alors plusieurs classes, chacune représentant un quartier, et nous voudrions alors retrouver ces classes dans l'espace de dimension inférieure. Dans ce cas, nous préférerons l'analyse factorielle discriminante.

Dans un premier temps, nous allons nous concentrer sur l'analyse en composantes principales. Nous détaillerons la méthode dans le cas classique pour passer à l'analyse en composantes principales symboliques. En effet, dans le cas symbolique, nous chercherons à nous ramener à une analyse en composantes principales classiques. Pour cela, nous devrons *transformer* nos données pour obtenir des données classiques. Cette transformation sera à la base des deux méthodes que nous allons envisager : la méthode des sommets et la méthode des centres. Nous finirons cette partie en illustrant la mise en oeuvre de ces méthodes par quelques exemples.

Nous passerons alors à l'analyse factorielle discriminante. De façon similaire, nous envisagerons le cas classique pour passer au cas symbolique. Dans cette méthode, nous chercherons également à transformer nos variables symboliques. Cette étape demandera la mise en place d'un système de codage.

Puisque nous voulons retrouver les classes auxquelles appartiennent les individus, nous aurons ici une étape supplémentaire : une fois les coordonnées des individus sur le plan factoriel connues, il nous faudra définir une règle qui nous permettra de les reclasser.

Nous terminerons par quelques applications à l'aide du logiciel d'analyse de données symboliques Sodas.

Les données

Comme nous l'avons signalé, les données peuvent être de deux types différents : classiques ou symboliques. Commençons par décrire ces données.

Les données classiques

Soit :

- $E = \{1, \ldots, n\}$ un ensemble de n objets
- $Y_1, \ldots, Y_p$, p variables que l'on mesure sur chacun de ces objets
- $\mathcal{Y}_1, \ldots, \mathcal{Y}_p$ les domaines de ces variables.

Nous avons alors :

$$\begin{aligned} Y_j : \; & E \to \mathcal{Y}_j \\ & k \to Y_j(k) = x_{kj} \end{aligned}$$

x_{kj} correspond à une seule et même valeur. Cette valeur est caractérisée par le type de variable classique auquel nous avons à faire. Elle sera :

- réelle dans le cas de variables *quantitatives*
- une catégorie dans le cas de variables *qualitatives* ou catégoriques.

Les variables quantitatives peuvent elles aussi être de 2 types différents :

- variables quantitatives *continues* qui prennent un nombre infini non dénombrable de valeurs dans $\mathbb{R}$
  Exemple : le chiffre d'affaire d'une entreprise

- variables quantitatives *discrètes* qui prennent un nombre fini ou infini dénombrable de valeurs dans $\mathbb{R}$
  Exemple : le nombre d'habitants dans une ville.

Ainsi que les variables qualitatives :

- variables qualitatives *nominales* où aucun ordre ne peut être défini sur les catégories de la variable
  Exemple : l'état civil d'une personne

- variables qualitatives *ordinales* où un ordre peut être défini
  Exemple : le grade obtenu à une session d'examen.

L'ensemble E

L'ensemble des objets E sur lequel nous mesurons nos variables peut être composé :

- d'individus k appelés objets du premier ordre ou
- de classes d'individus C_i appelées objets du second ordre.

Les objets du second ordre généralisent les objets du premier ordre. Par exemple, un objet du premier ordre peut être un homme et l'objet du second ordre correspondant, l'ensemble des hommes.

Des données classiques aux données symboliques

Les statistiques cherchent à étudier des faits, que nous appellerons *concepts*. Il arrive que ces concepts ne puissent pas être décrits par des données classiques car leur description nécessite des données plus complexes. Ces données seront appelées données symboliques.

Les objets symboliques

Les objets symboliques correspondent à la modélisation des concepts que nous souhaitons étudier. En effet, dans notre esprit, les concepts sont modélisés par une description du concept et un ensemble de caractéristiques nous permettant de reconnaître ce concept. Un objet symbolique correspond donc à un triplet :

$$(a, R, d)$$

où a est une fonction de reconnaissance, R une relation qui nous permet de reconnaître un objet observé en fonction de sa description et d est la description de l'objet que l'on a en tête, c'est-à-dire l'ensemble des caractéristiques qui nous permettent de reconnaître cet objet.

Les données symboliques

Dans le cas des données symboliques nous avons :

$$\begin{aligned} Y_j : \; & E \rightarrow \mathcal{Y}_j \\ & k \rightarrow Y(k) \end{aligned}$$

où $Y(k)$ est un ensemble de valeurs.

Comme dans le cas classique, les variables symboliques peuvent être de différents types :

- des variables *intervalles* : $Y(k)$ est alors un intervalle fermé borné de $\mathbb{R}$
  Exemple : la température d'une journée

- des variables *multivaluées* : $Y(k)$ est un sous-ensemble fini de $\mathcal{Y}_j$
  Exemple : la couleur d'une fleur

- des variables *modales* : $Y(k)$ est de la forme $(U(k), \pi_k)$ où $U(k)$ est un ensemble de catégories et π_k une distribution de fréquences associées à ces catégories.
  Exemple : les résultats d'un lancer de dé

On distingue deux types de variables multivaluées :

- les variables multivaluées catégoriques où $Y(k)$ a un nombre fini de catégories
- les variables multivaluées quantitatives où $Y(k)$ est un ensemble fini de nombres réels

Première partie

L'analyse en composantes principales

Chapitre 1

Analyse en composantes principales classique

1.1 Introduction

Comme nous venons de le voir, nous allons chercher à diminuer le nombre de variables intervenant dans la description des données. Pour cela, nous allons chercher des axes, que l'on appellera composantes principales, qui s'exprimeront comme une combinaison linéaire des variables de départ.

Ces axes devront tenir compte au maximum de la structure des données, c'est-à-dire de la dispersion des données. Pour cela, nous chercherons des composantes principales qui conserveront la plus grande proportion possible de la variance totale des données.

Dans un premier temps, nous considérerons des données quantitatives classiques c'est-à-dire que les variables sur lesquelles nous travaillerons auront pour domaine $\mathbb{R}$.

L'idée de l'analyse en composantes principales classique est de projeter l'ensemble des données sur un hyperplan de dimension inférieure à la dimension de l'espace de départ.
Cet hyperplan sera determiné par les vecteurs propres de la matrice de dispersion des données et les composantes principales seront construites à partir de ces vecteurs.

Dans ce chapitre, nous commencerons par représenter les données par une matrice X. Nous décrirons ensuite de façon détaillée les différentes étapes de l'analyse en composantes principales.

Il nous faudra ensuite définir quelques paramètres pour évaluer la qualité des résultats obtenus.

Nous terminerons ce chapitre par un exemple très simple pour illustrer les différents concepts mis en place et par un exemple plus concret afin de voir les interprétations que l'on peut faire des résultats d'une analyse en composantes principales.

1.2 Représentation des données

Considérons :

- n objets ou individus appartenant à un ensemble $\Omega = \{1, \ldots, n\}$. À chacun de ces objets, on associe un poids $p_i = \frac{1}{n}$; $i = 1, \ldots, n$

- $Y_1, \ldots, Y_p$ p variables *quantitatives* dont les domaines que nous noterons $\mathcal{Y}_j$ sont $\mathbb{R}$ $(j = 1, \ldots, p)$.

- n données $x_1, \ldots, x_n$ telles que $x_i \in \mathbb{R}^p$ $(i = 1, \ldots, n)$.

Chaque vecteur colonne $x_i = (x_{i1}, \ldots, x_{ip})'$ représente les différentes valeurs des variables Y_j pour l'objet i. En d'autres mots :

$$x_{ij} = Y_j(i).$$

L'ensemble des données peut être représenté par une matrice de données classiques X de dimension $n \times p$ dont les lignes sont les vecteurs x_i' :

$$X = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1' \\ \vdots \\ x_i' \\ \vdots \\ x_n' \end{pmatrix}.$$

1.3 La méthode

1.3.1 Idée générale

Nous allons considérer un hyperplan H^* de dimension s dans $\mathbb{R}^p$ $(s < p)$ sur lequel nous projetterons orthogonalement l'ensemble des points $x_1, \ldots, x_n$.

Ces projections sur H^* seront notées :

$$z_1^* = \pi_{H^*}(x_1), \ \ldots \ , z_n^* = \pi_{H^*}(x_n).$$

L'hyperplan H^* sera choisi de manière optimale parmi toutes les possibilités d'hyperplan H de dimension s dans l'espace de dimension p, c'est-à-dire en minimisant la distance entre les points et leurs projections :

$$Q(H) = \sum_{i=1}^{n} \| \ x_i - z_i \ \|^2$$

où $z_i = \pi_H(x_i)$.

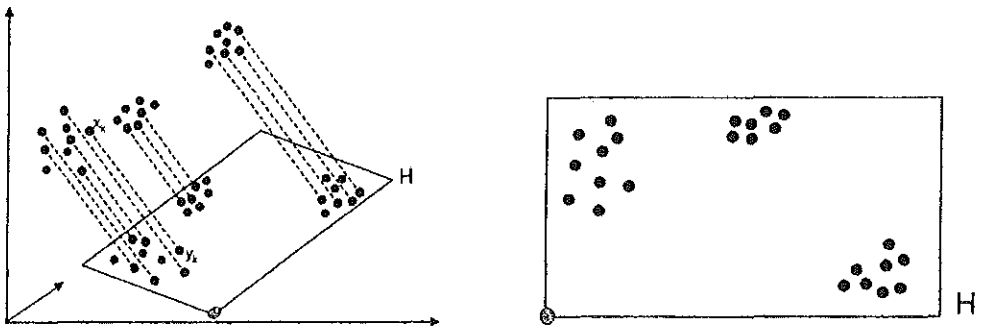


FIG. 1.1 – Projection des points sur le sous-espace (User manual for Sodas 2 software *http ://www.info.fundp.ac.be/asso*)

1.3.2 Première étape : Calcul du centroïde de l'ensemble des points

On détermine le centroïde de l'ensemble des points :

$$\bar{x} = (\bar{x}_1, \ldots, \bar{x}_j, \ldots, \bar{x}_p) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

où la j^e composante est donnée par la moyenne de la j^e colonne de X :

$$\bar{x_j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \; ; \quad j = 1, \ldots, p.$$

Par facilité dans des calculs ultérieurs, on soustrait cette moyenne à chacun des vecteurs de données, c'est-à-dire que l'on calcule les vecteurs :

$$\widetilde{x}_i = x_i - \bar{x}; \quad i = 1, \ldots, n.$$

On obtient ainsi des vecteurs de moyenne nulle c'est-à-dire que le centroïde des vecteurs $\widetilde{x}_i$ se trouve à l'origine.

On obtient alors une matrice centrée $\widetilde{X}$:

$$\widetilde{X} = \begin{pmatrix} \widetilde{x}_1 \\ \vdots \\ \widetilde{x}_i \\ \vdots \\ \widetilde{x}_n \end{pmatrix}.$$

1.3.3 Deuxième étape : Calcul de la matrice de dispersion

On calcule ensuite la matrice de dispersion S de dimension $p \times p$ définie par :

$$S = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' = \sum_{i=1}^{n} \widetilde{x}_i \, \widetilde{x}_i' = \widetilde{X}' \, \widetilde{X}.$$

Ou encore :

$$(S)_{rj} = \sum_{i=1}^{n}(x_{ir} - \bar{x_r})(x_{ij} - \bar{x_j}) = \sum_{i=1}^{n} \widetilde{x}_{ir}\, \widetilde{x}'_{ij}\ ; \quad j, r = 1, \ldots, p.$$

Nous pouvons voir que, en sommant les éléments de la diagonale principale de la matrice S, on obtient l'inertie totale de l'ensemble des points :

$$I_T = \sum_{i=1}^{n} \parallel x_i - \bar{x} \parallel^2 = \sum_{i=1}^{n} \parallel \widetilde{x}_i \parallel^2 .$$

1.3.4 Troisième étape : Calcul des valeurs propres et des vecteurs propres de S

On calcule ensuite les valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_p \geq 0$ de S ordonnées par valeur décroissante et les vecteurs propres *orthonormalisés* correspondants $v_1, \ldots, v_p \in \mathrm{R}^p$.

Les vecteurs propres nous permettent en effet de discerner la structure générale des données.

Par exemple, considérons deux variables Y_1 et Y_2, et représentons un ensemble de données sur un graphique ainsi que les vecteurs propres associés v_1 et v_2.

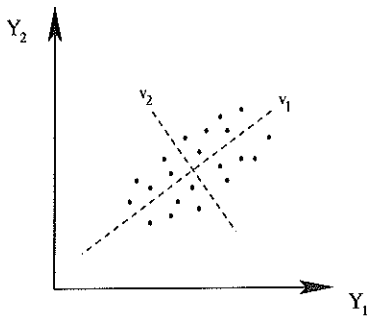


FIG. 1.2 – Les vecteurs propres d'un ensemble de données

Nous pouvons remarquer que le vecteur propre relatif à la plus grande valeur propre (v_1) indique la direction principale de dispersion des données.

Pour tenir compte au maximum de la dispersion des points, nous retiendrons donc les vecteurs propres correspondants aux valeurs propres les plus importantes.

1.3.5 Quatrième étape : Construction des composantes principales

Pour calculer les s composantes principales, on retient les s vecteurs propres de S qui correspondent aux s plus grandes valeurs propres, le j^e vecteur propre correspondant à la j^e plus grande valeur propre. Ce sont ces vecteurs propres qui déterminent l'hyperplan H^* que nous cherchons.

Puisque les composantes principales s'obtiennent par combinaison linéaire des variables initiales, la l^e composante principale Y_l^* est alors donnée par :

$$Y_l^* = v_l' Y \; ; \qquad l = 1, \ldots, s$$

où $Y = (Y_1, \ldots, Y_p)'$.

Remarques :

- Calculons la variance de la l^e composante principale :

$$Var(Y_l^*) = Var(v_l' Y) = v_l' S v_l.$$

Or, v_l est un vecteur propre orthonormé de S donc :

$$S v_l = \lambda_l v_l$$

et

$$Var(Y_l^*) = v_l' \lambda_l v_l = \lambda_l v_l' v_l = \lambda_l.$$

- La variance totale (ou inertie) des données de départ est :

$$I_T = \sum_{j=1}^{p} \lambda_j = trace(S)$$

- La l^e composante principale prend donc en compte la proportion :

$$t_l = \frac{\lambda_l}{trace(S)}$$

de la variance totale.

- Et les s composantes principales prennent en compte la proportion :

$$T = \frac{\sum_{l=1}^{s} \lambda_l}{trace(S)}$$

de la variance totale.

En considérant les vecteurs propres qui correspondent aux s plus grandes valeurs propres pour la construction des composantes principales, on tient compte de la plus grande variance possible.

Ce rapport T doit donc être le plus proche possible de 1.

1.3.6 Cinquième étape : Représentation des données dans l'espace de dimension s :

Les coordonnées des points dans l'espace de dimension s sont données par :

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{is} \end{pmatrix} = V_s' \widetilde{x}_i \; ; \qquad i = 1, \ldots, n$$

où :

$V_s = (v_1, \ldots, v_s)$ est la matrice de dimension $p \times s$ contenant les vecteurs propres de S et y_{il} est la valeur de la l^e composante principale pour l'objet i :

$$y_{il} = v_l' \, \widetilde{x}_i = \sum_{j=1}^{p} v_{lj} \, \widetilde{x}_{ij} \; ; \quad l = 1, \ldots s.$$

1.4 Choix du nombre de composantes principales

Pour évaluer le nombre s de composantes principales nécessaires pour avoir une bonne description des données, on regarde combien de valeurs propres ont une valeur significative. En effet, nous avons vu qu'une petite valeur propre ne nous apporte pas une grande information sur la variance des données.

Le critère le plus utilisé consiste à garder un nombre de composantes principales suffisant pour que T, la proportion de variance expliquée par les s premières composantes principales, ait une valeur élevée (0.8 ou 0.9). Citons d'autres critères possibles :

Le critère de Kaiser :

On considère autant de composantes principales que de vecteurs propres associés à des valeurs propres plus grandes que 1.

Le critère de Joliffe :

Ce critère est équivalent à celui de Kaiser mais ici, on ne retient que les valeurs propres dont la valeur est supérieure à 0.7.

Le critère de la valeur propre moyenne :

On considère autant de composantes principales que de vecteurs propres associés à des valeurs propres dont la valeur est supérieure à la moyenne des p valeurs propres.

La courbe de décroissance des valeurs propres :

On représente sur un graphique les différentes valeurs propres ordonnées par ordre décroissant. Le nombre de valeurs propres nécessaire à une bonne valeur de T sera déterminé par les points où la pente diminue fortement. Nous aurons donc autant de composantes principales que le nombre de valeurs propres indiquées par ce coude.

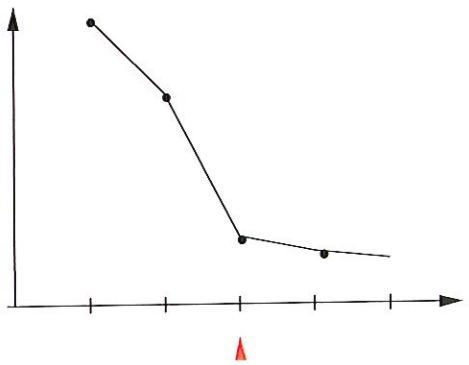


FIG. 1.3 – Courbe de décroissance des valeurs propres

1.5 Interprétation des résultats

Nous pouvons définir quelques paramètres pour estimer la qualité de représentation obtenue avec les composantes principales ainsi que la contribution de chaque composante principale à cette représentation.

* Le premier paramètre sera appelé la *contribution relative*, il mesure la qualité de représentation du vecteur x_i par rapport à la l^e composante principale :

$$COR(i, Y_l^*) = \frac{y_{il}^2}{\| \widetilde{x}_i \|^2}$$

où y_{il} est la valeur de la l^e composante principale pour l'objet i. Il s'agit en fait du carré du cosinus de l'angle formé par la j^e composante principale et le point x_i.

Si nous obtenons une valeur faible pour ce paramètre, nous devrons éviter d'interpréter la position de l'individu i sur le plan formé par les composantes principales.

* Un deuxième paramètre est appelé la *contribution absolue*, il mesure la contribution de x_i à la variance de la l^e composante principale :

$$CTR(i, Y_l^*) = \frac{p_i}{\lambda_l} y_{il}^2$$

où p_i est le poids de l'objet i.

Une valeur importante pour ce paramètre indique que l'individu est fort éloigné de l'origine.

* La contribution de x_i à la variance *totale* des données est :

$$INR(i) = \frac{p_i \parallel \widetilde{x}_i \parallel^2}{\sum_{j=1}^{p} \lambda_j} = \frac{p_i \parallel \widetilde{x}_i \parallel^2}{I_T}.$$

Remarque :

Les interprétations les plus significatives seront celles obtenues à partir d'une composante principale dont le pourcentage de variance expliquée est importante.

1.6 Exemples

1.6.1 Sur des données artificielles

Résumons les différentes étapes de l'analyse en composantes principales :

1. On calcule le centroïde des points et les vecteurs $\widetilde{x}_i$ pour $i = 1, \ldots, n$.
2. On calcule la matrice de dispersion S ainsi que ses valeurs propres et ses vecteurs propres.
3. On détermine le nombre de composantes principales nécessaires.
4. On construit ces composantes principales :

$$Y_l^* = v_l' Y.$$

5. On calcule les nouvelles coordonnées des points :

$$y_i = V_s' \widetilde{x}_i.$$

Soit 10 individus i ($\Omega = \{1, \ldots, 10\}$) sur lesquels on mesure 3 variables quantitatives Y_1, Y_2 et Y_3 :

	Y_1	Y_2	Y_3
1	1	1	2
2	0	4	1
3	2	1	1
4	3	2	1
5	2	3	0
6	0	2	2
7	4	1	2
8	3	0	1
9	4	3	0
10	1	3	0

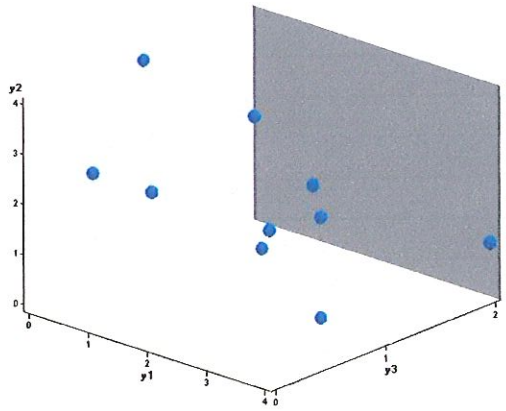


FIG. 1.4 – Représentation des individus dans l'espace formé par les 3 variables de départ

La matrice X de données classiques est :

$$X = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 4 & 1 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \\ 2 & 3 & 0 \\ 0 & 2 & 2 \\ 4 & 1 & 2 \\ 3 & 0 & 1 \\ 4 & 3 & 0 \\ 1 & 3 & 0 \end{pmatrix}.$$

On peut alors appliquer les différentes étapes de l'analyse en composantes principales :

Première étape : Calcul du centroïde des points

Le centroïde des points $\overline{x} = (\overline{x_1}, \overline{x_2}, \overline{x_3})$ est donné par :

$$\overline{x_1} = \frac{1}{10}\sum_{i=1}^{10} x_{i1} = 2$$

$$\overline{x_2} = \frac{1}{10}\sum_{i=1}^{10} x_{i2} = 2$$

$$\overline{x_3} = \frac{1}{10}\sum_{i=1}^{10} x_{i3} = 1.$$

On calcule ensuite les vecteurs $\widetilde{x}_i = (x_i - \overline{x})$ ce qui revient à calculer :

$$\widetilde{X} = \begin{pmatrix} \widetilde{x}_1 \\ \vdots \\ \widetilde{x}_{10} \end{pmatrix}$$

On obtient alors :

$$\widetilde{X} = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 4 & 1 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \\ 2 & 3 & 0 \\ 0 & 2 & 2 \\ 4 & 1 & 2 \\ 3 & 0 & 1 \\ 4 & 3 & 0 \\ 1 & 3 & 0 \end{pmatrix} - \begin{pmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 2 & 2 & 1 \\ 2 & 2 & 1 \\ 2 & 2 & 1 \\ 2 & 2 & 1 \\ 2 & 2 & 1 \\ 2 & 2 & 1 \\ 2 & 2 & 1 \\ 2 & 2 & 1 \end{pmatrix} = \begin{pmatrix} -1 & -1 & 1 \\ -2 & 2 & 0 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & -1 \\ -2 & 0 & 1 \\ 2 & -1 & 1 \\ 1 & -2 & 0 \\ 2 & 1 & -1 \\ -1 & 1 & -1 \end{pmatrix}$$

Deuxième étape : Calcul de la matrice de dispersion

Calculons la matrice de dispersion S qui sera de dimension 3×3 :

$$S = \sum_{i=1}^{10} (x_i - \overline{x})(x_i - \overline{x})'.$$

Ce qui revient à calculer :

$$S = \widetilde{X}' \, \widetilde{X}.$$

On a donc :

$$S = \begin{pmatrix} 2.2222 & -0.6667 & -0.2222 \\ -0.6667 & 1.5556 & -0.5556 \\ -0.2222 & -0.5556 & 0.6667 \end{pmatrix}.$$

On peut alors calculer les valeurs propres de S :

$$\lambda_1 = 0.2662$$
$$\lambda_2 = 1.5379$$
$$\lambda_3 = 2.6403$$

Les vecteurs propres correspondants sont :

$$v_1 = \begin{pmatrix} 0.8367 \\ -0.5444 \\ 0.059 \end{pmatrix}; \quad v_2 = \begin{pmatrix} 0.4807 \\ 0.6786 \\ -0.5553 \end{pmatrix}; \quad v_3 = \begin{pmatrix} 0.2623 \\ 0.4930 \\ 0.8295 \end{pmatrix}.$$

Troisième étape : Evaluation du nombre de composantes principales

On calcule la variance totale des données, ce qui revient à calculer l'inertie ainsi que le pourcentage t_i de variance expliqué par chaque valeur propre :

$$I = \sum_{i=1}^{3} \lambda_i = 4.4444$$

$$t_1 = \frac{\lambda_1}{40} = 0.5941$$
$$t_2 = \frac{\lambda_2}{40} = 0.3460$$
$$t_3 = \frac{\lambda_3}{40} = 0.0599$$

On décide alors de conserver 2 composantes principales (relatives au valeurs propres λ_1 et λ_2) qui prendront en compte la proportion :

$$T = \frac{\lambda_1 + \lambda_2}{40} = \frac{4.1782}{4.444} = 0.9401$$

de la variance totale des données.

Quatrième étape : Construction des composantes principales

Nos deux composantes principales sont alors données par :

$$* \; Y_1^* = v_1' \, Y = \begin{pmatrix} -0.8367 & -0.5444 & 0.059 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$$

$$\boxed{Y_1^* = -0.8367 \, Y_1 - 0.5444 \, Y_2 + 0.059 \, Y_3}$$

$$* \; Y_2^* = v_2' \, Y = \begin{pmatrix} 0.4807 & 0.6786 & -0.5553 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$$

$$\boxed{Y_2^* = 0.4807 \, Y_1 + 0.6786 \, Y_2 - 0.5553 \, Y_3}$$

Cinquième étape : Représentation des données dans l'espace de dimension 2 :

Les coordonnées du i^e objet dans l'espace des deux premières composantes principales sont alors :

$$y_i = V_2' \, (\widetilde{X})_i$$

où $\widetilde{X}_i$ est la i^e ligne de $\widetilde{X}$ et V_2' est la matrice contenant les vecteurs propres.

$$z_i^* = \begin{pmatrix} v_1' \\ v_2' \end{pmatrix} (\widetilde{X})_i = \begin{pmatrix} -0.8367 & -0.5444 & 0.0590 \\ 0.4807 & 0.6786 & -0.5553 \end{pmatrix} (\widetilde{X})_i$$

On obtient alors les coordonnées des 10 objets de départ dans l'espace à 2 dimensions formé par Y_1^* et Y_2^* :

Objets	Y_1^*	Y_2^*
1	-0.2332	-1.7147
2	-2.7623	0.3957
3	0.5444	-0.6786
4	0.8367	0.4807
5	-0.6035	1.2339
6	-1.6144	-1.5168
7	2.2769	-0.2724
8	1.9256	-0.8764
9	1.0699	2.1954
10	-1.4402	0.7532

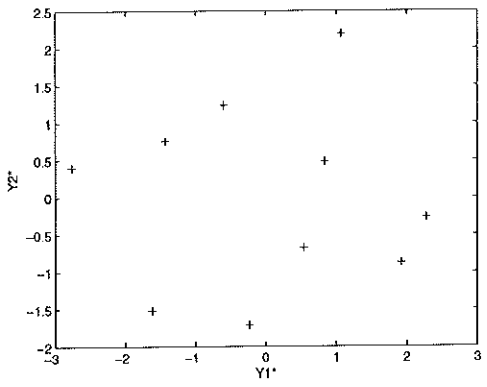


FIG. 1.5 – Représentation des individus dans le plan formé par les deux premières composantes principales

1.6.2 Sur des données réelles : Travail dans les différents pays d'Europe

FIG. 1.6 – L'Europe en 1979 (L. Swysen and G. Seret, Atlas Erasme, espace et société, 1996)

Présentation des données

Analysons le pourcentage de personnes travaillant dans différents secteurs dans les pays d'Europe en 1979.

Les différentes variables sont les suivantes :

- Y_1 Agriculture
- Y_2 Exploitation minière
- Y_3 Industries
- Y_4 Electricité
- Y_5 Construction
- Y_6 Industries de services
- Y_7 Finance
- Y_8 Social
- Y_9 Transport et communication

Ces données sont disponibles sur *http ://lib.stat.cmu.edu/DASL/Stories/EuropeanJobs.html.*

Pays	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9
Belgique	3.3	0.9	27.6	0.9	8.2	19.1	6.2	26.6	7.2
Danemark	9.2	0.1	21.8	0.6	8.3	14.6	6.5	32.2	7.1
France	10.8	0.8	27.5	0.9	8.9	16.8	6	22.6	5.7
Allemagne de l'Ouest	6.7	1.3	35.8	0.9	7.3	14.4	5	22.3	6.1
Irlande	23.2	1	20.7	1.3	7.5	16.8	2.8	20.8	6.1
Luxembourg	7.7	3.1	30.8	0.8	9.2	18.5	4.6	19.2	6.2
Pays-Bas	6.3	0.1	22.5	1	9.9	18	6.8	28.5	6.8
Royaume-Uni	2.7	1.4	30.2	1.4	6.9	16.9	5.7	28.3	6.4
Autriche	12.7	1.1	30.2	1.4	9	16.8	4.9	16.8	7
Finlande	13	0.4	25.9	1.3	7.4	14.7	5.5	24.3	7.6
Grèce	41.4	0.6	17.6	0.6	8.1	11.5	2.4	11	6.7
Norvège	9	0.5	22.4	0.8	8.6	16.9	4.7	27.6	9.4
Portugal	27.8	0.3	24.5	0.6	8.4	13.3	2.7	16.7	5.7
Espagne	22.9	0.8	28.5	0.7	11.5	9.7	8.5	11.8	5.5
Suède	6.1	0.4	25.9	0.8	7.2	14.4	6	32.4	6.8
Suisse	7.7	0.2	37.8	0.8	9.5	17.5	5.3	15.4	5.7
Turquie	66.8	0.7	7.9	0.1	2.8	5.2	1.1	11.9	3.2
Bulgarie	23.6	1.9	32.3	0.6	7.9	8	0.7	18.2	6.7
Tchécoslovaquie	16.5	2.9	35.5	1.2	8.7	9.2	0.9	17.9	7
Allemagne de l'Est	4.2	2.9	41.2	1.3	7.6	11.2	1.2	22.1	8.4
Hongrie	21.7	3.1	29.6	1.9	8.2	9.4	0.9	17.2	8
Pologne	31.1	2.5	25.7	0.9	8.4	7.5	0.9	16.1	6.9
Roumanie	34.7	2.1	30.1	0.6	8.7	5.9	1.3	11.7	5
URSS	23.7	1.4	25.8	0.6	9.2	6.1	0.5	23.6	9.3
Yougoslavie	48.7	1.5	16.8	1.1	4.9	6.4	11.3	5.3	4

Première étape : Calcul du centroïde des points

Calculons le centroïde des points $\overline{x} = (\overline{x}_1, \overline{x}_2, \overline{x}_3, \overline{x}_4, \overline{x}_5, \overline{x}_6, \overline{x}_7, \overline{x}_8, \overline{x}_9)$ où :

$$\overline{x}_j = \frac{1}{26} \sum_{i=1}^{26} x_{ij}$$

On a :

$$\overline{x}_1 = 19.1$$
$$\overline{x}_2 = 1.2$$
$$\overline{x}_3 = 27$$
$$\overline{x}_4 = 0.9$$
$$\overline{x}_5 = 8.2$$
$$\overline{x}_6 = 13$$
$$\overline{x}_7 = 4$$
$$\overline{x}_8 = 20$$
$$\overline{x}_9 = 6.5$$

$$\overline{x} = (19.1, 1.2, 27, 0.9, 8.2, 13, 4, 20, 6.5).$$

La matrice $\widetilde{X}$ est alors :

15.8308	-0.3538	0.5923	-0.0077	0.0346	6.1423	2.2	6.5769	0.6538
-9.9308	-1.1538	-5.2077	-0.3077	0.1346	1.6423	2.5	12.1769	0.5538
-8.3308	-0.4538	0.4923	-0.0077	0.7346	3.8423	2	2.5769	-0.8462
-12.4308	0.0462	8.7923	-0.0077	-0.8654	1.4423	1	2.2769	-0.4462
4.0692	-0.2538	-6.3077	0.3923	-0.6654	3.8423	-1.2	0.7769	-0.4462
-3.2308	-0.6538	0.5923	-0.4077	1.8346	5.1423	-2.4	0.0769	-0.8462
-11.4308	1.8462	3.7923	-0.1077	1.0346	5.5423	0.6	-0.8231	-0.3462
-12.8308	-1.1538	-4.5077	0.0923	1.7346	5.0423	2.8	8.4769	0.2538
-16.4308	0.1462	3.1923	0.4923	-1.2654	3.9423	1.7	8.2769	-0.1462
-6.4308	-0.1538	3.1923	0.4923	0.8346	3.8423	0.9	-3.2231	0.4538
-6.1308	-0.8538	-1.1077	0.3923	-0.7654	1.7423	1.5	4.2769	1.0538
22.2692	-0.6538	-9.4077	-0.3077	-0.0654	-1.4577	-1.6	-9.031	0.1538
-10.1308	-0.7538	-4.6077	-0.1077	0.4346	3.9423	0.7	7.5769	2.8538
8.6692	-0.9538	-2.5077	-0.3077	0.2346	0.3423	-1.3	-3.3231	-0.8462
3.7692	-0.4538	1.4923	-0.2077	3.3346	-3.2577	4.5	-8.2231	-1.0462
-13.0308	-0.8538	-1.1077	-0.1077	-0.9654	1.4423	2	12.3769	0.2538
-11.4308	-1.0538	10.7923	-0.1077	1.3346	4.5423	1.3	-4.6231	-0.8462
47.6692	-0.5538	-19.1077	-0.8077	-5.3654	-7.7577	-2.9	-8.1231	-3.3462

4.4692	0.6462	5.2923	-0.3077	-0.2654	-4.9577	-3.3	-1.8231	0.1538
-2.6308	1.6462	8.4923	0.2923	0.5346	-3.7577	-3.1	-2.1231	0.4538
-14.9308	1.6462	14.1923	0.3923	-0.5654	-1.7577	-2.8	2.0769	1.8538
2.5692	1.8462	2.5923	0.9923	0.0346	-3.5577	-3.1	-2.8231	1.4538
11.9692	1.2462	-1.3077	-0.0077	0.2346	-5.4577	-3.1	-3.9231	0.3538
15.5692	0.8462	3.0923	-0.3077	0.5346	-7.0577	-2.7	-8.3231	-1.5462
4.5692	0.1462	-1.2077	-0.3077	1.0346	-6.8577	-3.5	3.5769	2.7538
29.5692	0.2462	-10.2077	0.1923	-3.2654	-6.5577	7.3	-14.7231	-2.5462

Deuxième étape : Calcul de la matrice de dispersion

Nous pouvons alors calculer S qui sera ici de dimension 9×9 :

$$S = \widetilde{X}' \, \widetilde{X}.$$

Troisième étape : Calcul des valeurs propres et des vecteurs propres de S

Les valeurs propres de S sont :

$$\begin{aligned}
\lambda_1 &= 303.4581 \\
\lambda_2 &= 43.7017 \\
\lambda_3 &= 15.2074 \\
\lambda_4 &= 5.6394 \\
\lambda_5 &= 2.4434 \\
\lambda_6 &= 1.0460 \\
\lambda_7 &= 0.4208 \\
\lambda_8 &= 0.0649 \\
\lambda_9 &= 0.0019.
\end{aligned}$$

L'inertie totale I est de 9299.6 et les pourcentages de variance expliquée par chaque valeur propre sont :

$$
\begin{aligned}
t_1 &= 0.8158 \\
t_2 &= 0.1175 \\
t_3 &= 0.0409 \\
t_4 &= 0.0152 \\
t_5 &= 0.0066 \\
t_6 &= 0.0028 \\
t_7 &= 0.0011 \\
t_8 &= 1.7454.10^{-4} \\
t_9 &= 5.14.10^{-6}.
\end{aligned}
$$

On décide alors de prendre 2 composantes principales.

Les vecteurs propres correspondant aux 2 premières valeurs propres sont :

$$
v_1 = \begin{pmatrix} 0.8918 \\ 0.0019 \\ -0.2713 \\ -0.0084 \\ -0.0496 \\ -0.1918 \\ -0.0311 \\ -0.298 \\ -0.0454 \end{pmatrix} ; \quad v_2 = \begin{pmatrix} -0.0068 \\ 0.0923 \\ 0.7703 \\ 0.012 \\ 0.069 \\ -0.2344 \\ -0.1301 \\ -0.5668 \\ -0.0099 \end{pmatrix}.
$$

Quatrième étape : Construction des composantes principales

Nos deux composantes principales sont données par :

$$Y_1^* = v_1' \, Y = \begin{pmatrix} 0.8918 \\ 0.0019 \\ -0.2713 \\ -0.0084 \\ -0.0496 \\ -0.1918 \\ -0.0311 \\ -0.2980 \\ -0.0454 \end{pmatrix}' \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \end{pmatrix}$$

$$\begin{aligned} Y_1^* =& 0.8918Y_1 + 0.0019Y_2 - 0.2713Y_3 - 0.0084Y_4 - 0.0496Y_5 - 0.1918Y_6 - 0.0311Y_7 \\ & - 0.298Y_8 - 0.0454Y_9. \end{aligned}$$

$$Y_2^* = v_2' \, Y = \begin{pmatrix} -0.0068 \\ 0.0923 \\ 0.7703 \\ 0.012 \\ 0.069 \\ -0.2344 \\ -0.1301 \\ -0.5668 \\ -0.0099 \end{pmatrix}' \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \end{pmatrix}$$

$$\begin{aligned} Y_2^* =& -0.0068Y_1 + 0.0923Y_2 + 0.7703Y_3 + 0.012Y_4 + 0.069Y_5 - 0.2344Y_6 - 0.1301Y_7 \\ & - 0.5668Y_8 - 0.0099Y_9. \end{aligned}$$

Cinquième étape : Représentation des données dans l'espace de dimension 2 :

Les coordonnées du i^e objet dans l'espace à deux dimensions sont alors calculées par :

$$z_i^* = V_2' \, (X^*)_i$$

où $(X^*)_i$ est la i^e ligne de X^* et V_2' est la matrice contenant les vecteurs propres de S correspondants aux 2 plus grandes valeurs propres.

On obtient alors les résultats suivants pour les 26 pays :

	Belgique	Danemark	France	Allemagne de l'Ouest	Irlande	Italie	Luxembourg
Y_1^*	-17.5167	-11.4967	-9.1287	-14.3934	4.4582	-4.0267	-12.0898
Y_2^*	-4.9262	-11.6618	-2.1683	5.0475	-6.1316	-0.3889	2.3324

	Pays-Bas	Royaume-Uni	Autriche	Finlande	Grèce	Norvège	Portugal
Y_1^*	-13.9005	-18.7287	-6.4714	-6.837	25.4271	-10.972	9.4039
Y_2^*	-9.7236	-3.3318	3.3566	-3.9763	-1.8047	8.8588	-0.0857

	Espagne	Suède	Suisse	Turquie	Bulgarie	Tchéco-slovaquie	Allemagne de l'Est	Hongrie
Y_1^*	5.775	-15.312	-12.6838	52.1156	4.1568	-3.2461	-17.4155	3.1357
Y_2^*	6.1587	-8.5267	9.7792	-8.6417	6.7069	9.2347	10.7323	4.987

	Pologne	Roumanie	URSS	Yougoslavie
Y_1^*	13.3157	17.0113	4.587	34.8326
Y_2^*	2.9448	9.1252	-0.872	0.6927

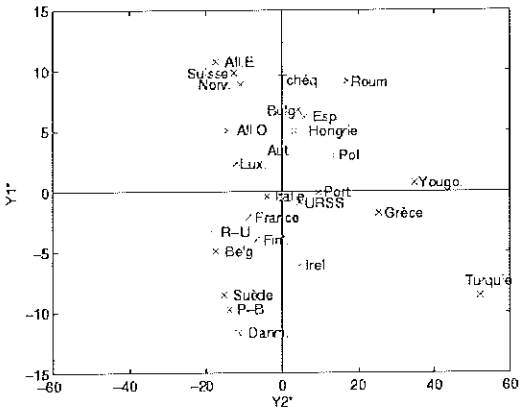


FIG. 1.7 – Représentation des pays dans l'espace formé par les deux premières composantes principales

Interprétation des résultats :

Nous pouvons calculer les différents paramètres d'interprétation.

	$COR_I^1(i, Y_1^I)$	$COR_I^1(i, Y_2^I)$	$CTR_I(i, Y_1^I)$	$CTR_I(i, Y_2^I)$	$INR_I(i)$
Belgique	0.9096	0.0719	0.0016	$8.5431.10^{-4}$	0.0014
Danemark	0.4642	0.4777	$6.7009.10^{-4}$	0.0048	0.0012
France	0.8635	0.0487	$4.2248.10^{-4}$	$1.6551.10^{-4}$	$3.9915.10^{-4}$
Allemagne de l'Ouest	0.8598	0.1057	0.0011	$8.9689.10^{-4}$	$9.9692.10^{-4}$
Irlande	0.2685	0.508	$1.0076.10^{-4}$	0.0013	$3.061.10^{-4}$
Italie	0.3431	0.0032	$8.2202.10^{-5}$	$5.3242.10^{-6}$	$1.9717.10^{-4}$
Luxembourg	0.8057	0.03	$7.4101.10^{-4}$	$1.9151.10^{-4}$	$7.5028.10^{-4}$
Pays-Bas	0.6561	0.3211	$9.7959.10^{-4}$	0.0033	0.0012
Royaume-Uni	0.9506	0.0301	0.0018	$3.9079.10^{-4}$	0.0015
Autriche	0.5323	0.1432	2.1232	$3.9664.10^{-4}$	$3.2539.10^{-4}$
Finlande	0.7195	0.2434	$2.3699.10^{-4}$	$5.5662.10^{-4}$	$2.687.10^{-4}$
Grèce	0.9634	0.0049	0.0033	$1.1465.10^{-4}$	0.0028

	$COR_I^1(i,Y_1^I)$	$COR_I^1(i,Y_2^I)$	$CTR_I(i,Y_1^I)$	$CTR_I(i,Y_2^I)$	$INR_I(i)$
Norvège	0.5838	0.3806	$6.1032.10^{-4}$	0.0028	$8.5288.10^{-4}$
Portugal	0.9205	$7.6451.10^{-5}$	$4.4833.10^{-4}$	$2.5856.10^{-7}$	$3.9733.10^{-4}$
Espagne	0.2618	0.2978	$1.6908.10^{-4}$	0.0013	$5.2682.10^{-4}$
Suède	0.7061	0.219	0.0012	0.0026	0.0014
Suisse	0.5464	0.3248	$8.1562.10^{-4}$	0.0034	0.0012
Turquie	0.9655	0.0265	0.0138	0.0026	0.0116
Bulgarie	0.1977	0.5148	$8.76.10^{-5}$	0.0016	$3.6139.10^{-4}$
Tchéco-slovaquie	0.0953	0.7713	$5.3422.10^{-5}$	0.003	$4.5728.10^{-4}$
Allemagne de l'Est	0.6797	0.2581	0.0015	0.0041	0.0018
Hongrie	0.1964	0.4967	$4.985.10^{-5}$	$8.7551.10^{-4}$	$2.0706.10^{-4}$
Pologne	0.88	0.043	$8.9891.10^{-4}$	$3.0529.10^{-4}$	$8.3334.10^{-4}$
Roumanie	0.7579	0.2181	0.0015	0.0029	0.0016
URSS	0.2039	0.0074	$1.0667.10^{-4}$	$2.6767.10^{-5}$	$4.2673.10^{-4}$
Yougoslavie	0.927	$3.6666.10^{-4}$	0.0062	$1.6894.10^{-5}$	0.0054

Nous pouvons également calculer les indices de corrélations entre les variables initiales $Y_1, \ldots, Y_9$ et les composantes principales Y_1^* et Y_2^* que nous avons obtenues :

	Y_1^*	Y_2^*
Y_1	0.9992	-0.0029
Y_2	0.0345	0.6293
Y_3	-0.6743	0.7266
Y_4	-0.3884	0.2111
Y_5	-0.525	0.2771
Y_6	-0.7303	-0.3387
Y_7	-0.1932	-0.3064
Y_8	-0.7602	-0.5486
Y_9	-0.5679	-0.0470

Nous pouvons alors voir que la première composante principale est principalement corrélée avec la variable Y_1 représentant l'agriculture. Elle peut donc être vue comme distinguant les pays à tendance agricole ($Y_1^* > 0$) et les pays plus industriels ($Y_1^* < 0$).

La deuxième composante principale est quant à elle principalement corrélée avec les variables Y_2 (les mines), Y_3 (les industries) et Y_8 (le secteur social).

Nous pouvons donc conclure que les pays fortement industrialisés (également dans l'industrie minière) et agricoles auront une valeur positive pour les deux composantes principales. Les pays ayant une première et une deuxième composante principale négatives seront les pays industriels où le secteur social est bien développé. Les pays qui ont une première composante principale positive et une seconde négative seront les pays principalement agricoles où le secteur social est développé. Les pays qui ont une première composante principale négative et une deuxième positive seront les pays industriels (y compris l'industrie minière) où le secteur social n'est pas développé.

Trois pays se distinguent des autres :

- La Grèce et la Turquie qui ont leur première composante principale positive et la seconde négative (surtout la Turquie) ce qui signifie que ces pays sont principalement agricoles, que le secteur social y est développé.
- La Yougoslavie avec les deux composantes principales positives (surtout la première) ce qui signifie que c'est un pays agricole mais où les industries sont un minimum développées ainsi que les mines.

Chapitre 2

Analyse en composantes principales pour des données intervalles

2.1 Introduction

Dans ce chapitre, nous allons chercher à généraliser l'analyse en composantes principales classique. La méthode que nous allons développer ici ne sera valable que pour des données de type intervalle. C'est-à-dire que nous travaillerons avec des variables qui auront pour domaine $\mathcal{I}$, l'ensemble des intervalles fermés bornés de $\mathbb{R}$.

Nous aborderons deux méthodes de généralisation :

- La méthode des sommets
- La méthode des centres.

Ces deux méthodes se différencient par la façon dont les données sont représentées : par des sommets ou par des centres comme l'indique le nom des méthodes. Nous verrons que la méthode des centres est préférable lorsque nous avons un grand nombre de variables.

Dans les deux cas, nous nous ramènerons à une application de l'analyse en composantes principales classique.

Nous chercherons également à généraliser les paramètres d'interprétation de la méthode classique et nous terminerons par un exemple simple sur des données artificielles et un exemple de données réelles.

Lorsque nous avons des données symboliques autres que des données intervalles, la généralisation de l'analyse en composantes principales classique se fait par une analyse canonique généralisée symbolique que nous aborderons dans le cadre de l'analyse factorielle discriminante.

2.2 Représentation des données de type intervalle

De manière similaire au cas classique, nous partons avec :

- n objets que nous noterons $u \in \Omega = \{1, \ldots, n\}$
- $Y_1, \ldots, Y_p$, p variables intervalles dont les domaines $\mathcal{Y}_j$ $(j = 1, \ldots, p)$ sont $\mathcal{I}$, l'ensemble des intervalles fermés bornés de $\mathbb{R}$
- n données intervalles $x_1, \ldots, x_n$ où $x_i = (x_{i1}, \ldots, x_{ip})' \in \mathbb{R}^p$ avec $x_{ij} = Y_j(x_i) = [\underline{x}_{ij}, \overline{x}_{ij}]$.
  Ces données x_i seront donc de la forme :

$$x_i' = ([\underline{x_{i1}}, \overline{x_{i1}}], \ldots, [\underline{x_{ip}}, \overline{x_{ip}}]).$$

Nous pouvons également représenter ces données par une matrice de données intervalles :

$$\underline{X} = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix}$$

où l'élément x_{ij} représente l'intervalle $[\underline{x_{ij}}, \overline{x_{ij}}] \in \mathcal{I}$.

Soit $x_i' = (x_{i1}, \ldots, x_{ip}) = ([\underline{x_{i1}}, \overline{x_{i1}}], \ldots, [\underline{x_{ip}}, \overline{x_{ip}}])$ le vecteur de données symboliques relatif à l'objet i et soit q_i le nombre d'intervalles x_{ij} non triviaux.

Remarque :

Un intervalle $[\underline{x_{ij}}, \overline{x_{ij}}]$ est dit *trivial* si $\underline{x_{ij}} = \overline{x_{ij}}$ c'est-à-dire si il ne contient qu'une seule valeur.

Dans un espace de dimension p, l'objet x_i sera représenté par un hyper-rectangle R_i à 2^{q_i} sommets où q_i est le nombre de variables intervalles intervenant dans la description de l'individu i et dont les longueurs des côtés sont données par les longueurs des différents intervalles x_{ij} intervenant dans la description de l'objet i.

Par exemple, si on prend $p = 2$, l'intervalle $x'_i = ([\underline{x_{i1}}, \overline{x_{i1}}], [\underline{x_{i2}}, \overline{x_{i2}}])$ peut être représenté par le rectangle suivant où les différents sommets sont :

$$(\underline{x}_{i1}, \underline{x}_{i2})$$
$$(\underline{x}_{i1}, \overline{x}_{i2})$$
$$(\overline{x}_{i1}, \underline{x}_{i2})$$
$$(\overline{x}_{i1}, \overline{x}_{i2})$$

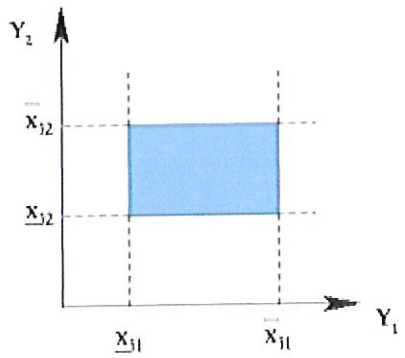


FIG. 2.1 – Représentation d'un individu décrit par deux variables intervalles

Prenons le cas où 7 objets notés i ($i = 1, \ldots, 7$) sont décrits par 3 variables intervalles.

Nous pouvons alors voir sur la figure suivante que, suivant la valeur de q_i, un hyperrectangle peut aussi désigner :

- un point si $q_i = 0$ comme l'objet 7
- un segment lorsque $q_i = 1$ comme l'objet 6
- ou un rectangle si $q_i = 2$ comme l'objet 5.

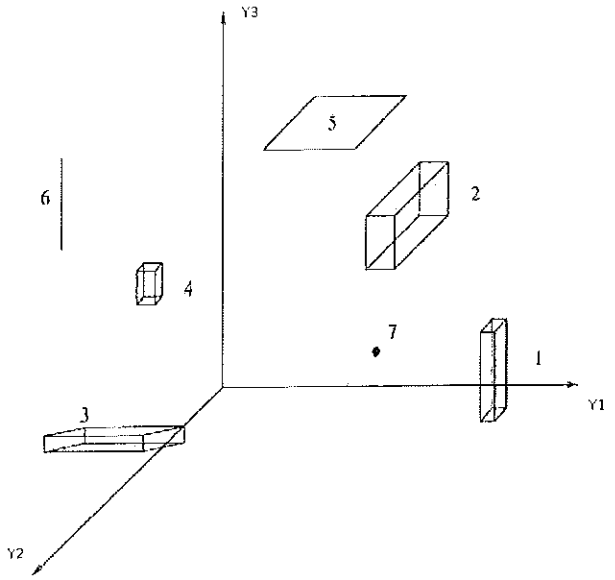


FIG. 2.2 – Différentes représentations d'individus décrits par des variables intervalles

2.3 Pondération des individus

Comme pour l'analyse en composantes principales classique, à chaque individu i, nous associons un poids p_i. Cette pondération peut s'effectuer de différentes manières.

La plus fréquente, et celle que nous adopterons ici, consiste à attribuer aux objets i des poids identiques :

$$p_i = \frac{1}{n} \; ; \quad i = 1, \ldots, n.$$

Nous pourrions également envisager un système de pondération dans lequel nous attribuons à chaque objet un poids p_i proportionnel à l'amplitude de la variation représentée par son intervalle, c'est-à-dire qu'un objet aura un poids plus important si le volume de l'hyperrectangle R_i est plus important ou encore si la précision des intervalles est plus faible.

$$p_i = \frac{V_i}{\sum_{i=1}^{n} V_i} \; ; \quad i = 1, \ldots, n$$

où V_i est le volume de l'hyperrectangle R_i représentant l'individu i.

Un dernier système de pondération consiste, à l'opposé du second, à attribuer des poids importants aux objets représentés par des hyperrectangles de faible volume, c'est-à-dire qu'un objet aura plus de poids si la précision des intervalles intervenant dans sa description est plus importante :

$$p_i = \frac{1 - \frac{V_i}{\sum V_i}}{\sum_{i=1}^{n} \left(1 - \frac{V_i}{\sum V_i}\right)} \quad i = 1, \ldots, n.$$

2.4 La méthode des sommets

2.4.1 Idée générale

Nous allons représenter chaque donnée x_i par un hyperrectangle R_i à 2^{q_i} sommets de $\mathbb{R}^p$.

L'idée de la méthode des sommets est d'appliquer l'analyse en composantes principales classique à l'ensemble des sommets appartenant aux n individus.

Les composantes principales intervalles seront alors construites à partir des composantes principales classiques obtenues sur l'ensemble des sommets.

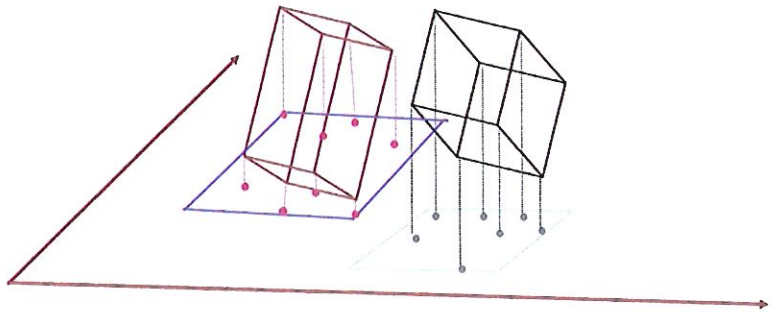


FIG. 2.3 – Projection des individus sur le sous-espace formé par les composantes principales (R. Verde, Analyse des données symboliques, Université de Namur, 2006)

2.4.2 Première étape : Représentation des données

Chaque hyperrectangle R_i peut aussi être représenté par une matrice M_i de dimension $2^{q_i} \times p$ dont chaque ligne contient les coordonnées d'un sommet de l'hyperrectangle.

Exemples :

Pour l'individu $1 = ([\underline{x}_{11}, \overline{x}_{11}], [\underline{x}_{12}, \overline{x}_{12}], [\underline{x}_{13}, \overline{x}_{13}])$:

$$M_1 = \begin{pmatrix} \underline{x}_{11} & \underline{x}_{12} & \underline{x}_{13} \\ \underline{x}_{11} & \underline{x}_{12} & \overline{x}_{13} \\ \underline{x}_{11} & \overline{x}_{12} & \underline{x}_{13} \\ \underline{x}_{11} & \overline{x}_{12} & \overline{x}_{13} \\ \overline{x}_{11} & \underline{x}_{12} & \underline{x}_{13} \\ \overline{x}_{11} & \underline{x}_{12} & \overline{x}_{13} \\ \overline{x}_{11} & \overline{x}_{12} & \underline{x}_{13} \\ \overline{x}_{11} & \overline{x}_{12} & \overline{x}_{13} \end{pmatrix}.$$

Pour l'individu $5 = ([\underline{x}_{51}, \overline{x}_{51}], [\underline{x}_{52}, \overline{x}_{52}], [x_{53}, x_{53}])$:

$$M_5 = \begin{pmatrix} \underline{x}_{51} & \underline{x}_{52} & x_{53} \\ \underline{x}_{51} & \overline{x}_{52} & x_{53} \\ \overline{x}_{51} & \underline{x}_{52} & x_{53} \\ \overline{x}_{51} & \overline{x}_{52} & x_{53} \end{pmatrix}.$$

Pour l'individu $6 = ([x_{61}, x_{61}], [x_{62}, x_{62}], [\underline{x}_{63}, \overline{x}_{63}])$:

$$M_6 = \begin{pmatrix} x_{61} & x_{62} & \underline{x}_{63} \\ x_{61} & x_{62} & \overline{x}_{63} \end{pmatrix}.$$

Pour l'individu $7 = ([x_{71}, x_{71}], [x_{72}, x_{72}], [x_{73}, x_{73}])$:

$$M_7 = \begin{pmatrix} x_{71} & x_{72} & x_{73} \end{pmatrix}.$$

Remarque :

Dans chaque colonne j, on retrouve 2^{q_i-1} fois les 2 bornes de l'intervalle x_{ij} qui correspondent aux valeurs prises par la j^e variable sur le i^e objet. Pour retrouver les bornes de l'intervalle j à partir de la matrice M_i, il suffit donc de prendre le minimum et le maximum de la j^e colonne.

2.4.3 Deuxième étape : Construction de la matrice M

On regroupe toutes les matrices M_i dans une nouvelle matrice M de dimension $\sum_{i=1}^{n} 2^{q_i} \times p$:

$$M = \begin{pmatrix} M_1 \\ \vdots \\ M_n \end{pmatrix}.$$

Cette matrice contient donc les coordonnées de tous les sommets appartenant aux différents objets.

2.4.4 Troisième étape : Application de l'analyse en composantes principales classique

On applique l'analyse en composantes principales classique à la matrice M c'est-à-dire que l'on va chercher à réduire le nombre de coordonnées de chaque sommet.

Soit $Y_1^*, \ldots, Y_s^*$ les s premières composantes principales obtenues par l'analyse en composantes principales classique et $\lambda_1 \geq \ldots \geq \lambda_s \geq 0$ les valeurs propres associées.

2.4.5 Quatrième étape : Construction des composantes principales intervalles

On construit alors les s composantes principales de type intervalle $Y_1^I, \ldots, Y_s^I$ à partir des s composantes principales classiques $Y_1^*, \ldots, Y_s^*$.

Pour cela considérons :

- L_i l'ensemble des indices k de lignes de la matrice M qui se réfèrent aux sommets de R_i.

- y_{kl} est la valeur de la composante principale classique Y_l^* pour le sommet de R_i dont l'indice de ligne est k.

Pour chaque ensemble L_i, on construit ainsi une matrice M_i^* de dimension $2^{q_i} \times s$ dont les lignes ($\in L_i$) représentent les coordonnées des sommets de l'hyperrectangle R_i dans un espace de dimension s.

- Comme nous l'avons vu, pour retrouver les bornes de l'intervalle correspondant aux valeurs prises par la j^e variable, il suffit de prendre la valeur minimale et la valeur maximale de la j^e colonne de la matrice représentant les coordonnées des sommets de R_i c'est-à-dire de la matrice M_i^*.

  La valeur de la l^e composante principale de type intervalle Y_l^I pour le i^e objet est alors :

$$\boxed{y_{il} = [\underline{y_{il}}, \overline{y_{il}}]}$$

où

$$\underline{y_{il}} = \min_{k \in L_i} y_{kl}$$

$$\overline{y_{il}} = \max_{k \in L_i} y_{kl}.$$

2.4.6 Interprétation des résultats

On peut généraliser les paramètres d'interprétation obtenus pour la méthode classique au cas des données intervalles.

* La qualité de représentation du vecteur x_i par rapport à la l^e composante principale Y_l^I est donnée par :

$$COR_I^1(i, Y_l^I) = \frac{\sum_{k \in L_i} \omega_k \, y_{kl}^2}{\sum_{k \in L_i} q_k \, d^2(k, G)}$$

où $G \in R^p$ est le centroïde des $n.2^p$ lignes de la matrice M et $d(k, G)$ est la distance (euclidienne) entre la ligne $k \in L_i$ de la matrice M et G c'est-à-dire la distance entre le sommet k de R_i et G, ω_k est le poids du sommet k de L_i

$\sum_{k \in L_i} d^2(k,G) = \sum_{k \in L_i} \| x_k - G \|^2$ est l'inertie totale des sommets de R_i.

Les poids ω_k étant identiques pour les sommets d'un même hyperrectangle R_i dont l'ensemble des indices de ligne dans la matrice est L_i, nous avons :

$$COR_I^1(i, Y_l^I) = \frac{\sum_{k \in L_i} y_{kl}^2}{\sum_{k \in L_i} d^2(k,G)}.$$

Ce paramètre indique le rapport entre la contribution de l'ensemble des sommets de L_i à l'inertie λ_j de la j^e composante principale et la contribution de l'ensemble des sommets à l'inertie totale.

Nous avons en effet que :

$$\begin{aligned} \lambda_l &= var(Y_l^*) \\ &= \sum_{i=1}^{n} \sum_{k \in L_i} y'_{kl}\, y_{kl} \\ &= \sum_{i=1}^{n} \sum_{k \in L_i} y_{kl}^2. \end{aligned}$$

$*$ Nous pouvons définir un deuxième paramètre pour évaluer la qualité de représentation du vecteur x_i :

$$COR_I^2(i, Y_j^I) = \frac{1}{2^{q_i}} \sum_{k \in L_i} \frac{y_{kj}^2}{d^2(k,G)}$$

qui indique le cosinus moyen des angles entre chacun des 2^{q_i} sommets de R_i et la j^e composante principale.

Remarque :

Le premier paramètre nous donnera des contributions plus faibles que le second puisque :

$$\frac{a+c}{b+d} \leq \frac{a}{b} + \frac{c}{d}$$

où $a, b, c, d \in \mathbb{R}^+$ et donc :

$$\frac{y_{1l}^2 + \ldots + y_{kl}^2}{d^2(1,G) + \ldots + d^2(k,G)} \leq \frac{y_{1l}^2}{d^2(1,G)} + \ldots + \frac{y_{kl}^2}{d^2(k,G)}.$$

$*$ La contribution de x_i à la variance λ_l de la l^e composante principale est :

$$CTR_I(i, Y_l^I) = \frac{\sum_{k \in L_i} \omega_k \, y_{kl}^2}{\lambda_l}.$$

Or $\omega_k = p_i/2^{q_i}$ est le poids du sommet k de R_i (p_i étant le poids de l'objet i) donc :

$$CTR_I(i, Y_l^I) = \frac{p_i}{2^{q_i}.\lambda_l} \sum_{k \in L_i} y_{kl}^2.$$

$*$ La contribution de x_i au SST des $n.2^{q_i}$ sommets représentant les n hyperrectangles est donnée par :

$$INR_I(i) = \frac{\sum_{k \in L_i} \omega_k.d^2(k,G)}{I_T} = \frac{p_i}{2^p} \frac{\sum_{k \in L_i} d^2(k,G)}{I_T}$$

où $I_T = \sum_{j=1}^{p} \lambda_j$ est la variance totale du nuage des $n.2^{q_i}$ sommets dans l'espace de dimension p.

Ces deux dernières contributions reviennent à effectuer des moyennes pondérées $(p_i/2^p)$ des contributions des sommets de L_i associés à l'objet i à l'inertie λ_l du l^e axe factoriel et à l'inertie totale respectivement.

2.4.7 Lien avec la méthode classique

Nous pouvons montrer que l'analyse en composantes principales classique est un cas particulier de la méthode des sommets.

Les données classiques sont en effet un cas particulier des données intervalles : ce sont des données décrites par p intervalles triviaux :

$$\underline{x_{i1}} = \overline{x_{i1}} = x_{i1}, \ \ldots, \ \underline{x_{ip}} = \overline{x_{ip}} = x_{ip}.$$

Les hyperrectangles R_i sont alors des points de coordonnées $(x_{i1}, \ldots, x_{ip})$, les matrices M_i sont des matrices lignes et la matrice M est en fait la matrice X des données classiques.
On doit donc appliquer l'analyse en composantes principales classique à la matrice X.

Il ne nous reste alors plus qu'à montrer que la construction des composantes principales intervalles ne modifie pas les résultats obtenus par la méthode classique.

- L'ensemble L_i ne contient qu'un seul indice de ligne : i.
- On a une seule valeur pour y_{kl} et donc :

$$\underline{y_{il}} = \min_{k \in L_i} y_{kl} = y_{il} = \max_{k \in L_i} y_{kl} = \overline{y_{il}}.$$

- On a alors que y_{il}, la valeur de la composante principale intervalle Y_l^I pour l'objet i, est la valeur de la composante principale classique Y_l^* pour l'objet i.

On peut donc voir que cette dernière étape ne modifie pas les valeurs obtenues avec la méthode classique et donc que l'analyse en composantes principales classique est bien un cas particulier de la méthode des sommets.

2.5 La méthode des centres

2.5.1 Idée générale

Dans cette méthode, nous caractériserons chaque hyperrectangle R_i par son centre que l'on notera c_i.

De nouveau, ces centres auront p composantes réelles et nous pourrons donc leur appliquer l'analyse en composantes principales classique. Les composantes principales intervalles seront alors construites à partir des résultats obtenus sur l'ensemble des centres.

Chaque centre $c_i = (x_{i1}^c, \ldots, x_{ip}^c)$ a pour coordonnées :

$$x_{ij}^c = \frac{\underline{x_{ij}} + \overline{x_{ij}}}{2} \; ; \quad j = 1, \ldots, p; \quad i = 1, \ldots, n.$$

2.5.2 Première étape : Construction de $\widetilde{X}$

On calcule les centres des n hyperrectangles pour construire la matrice $\widetilde{X}$ définie par :

$$\widetilde{X} = \begin{pmatrix} x_{11}^c & \ldots & x_{1p}^c \\ \vdots & \vdots & \vdots \\ x_{n1}^c & \ldots & x_{np}^c \end{pmatrix}$$

où la ligne i représente les coordonnées du centre c_i de R_i.

Les colonnes de $\widetilde{X}$ représentent les nouvelles valeurs des variables $Y_1, \ldots, Y_p$, mesurées sur les différents sommets et sont notées : $Y_1^c, \ldots, Y_p^c$.

2.5.3 Deuxième étape : Application de l'analyse en composantes principales classique

On applique l'analyse en composantes principales à la matrice $\widetilde{X}$ c'est-à-dire aux centres des n hyperrectangles.

Soit y_{il}^c la valeur de la l^e composante principale pour c_i.

2.5.4 Troisième étape : Construction des composantes principales intervalles

Nous pouvons alors déterminer les valeurs des composantes principales intervalles pour chaque objet i.

Soit :

$$\overline{x_j^c} = \frac{\sum_{i=1}^{n} x_{ij}^c}{n}$$

la moyenne de la variable Y_j^c c'est-à-dire la moyenne des éléments de la j^e colonne de $\widetilde{X}$.

Par définition dans la méthode classique, la l^e composante principale du centre c_i est donnée par :

$$y_{il}^c = \sum_{j=1}^{p} (x_{ij} - \overline{x_j^c})\, v_{jl}$$

où $v_l = (v_{1l}, \ldots, v_{pl})$ est le l^e vecteur propre de S.

Puisque les coordonnées x_{ij}^c du i^e centre sont localisées entre $\underline{x_{ij}}$ et $\overline{x_{ij}}$, nous pouvons trouver un intervalle $[\underline{y_{ij}}, \overline{y_{ij}}]$ dans lequel les valeurs possibles de la l^e composante principale y_{il}^c doivent se trouver.

Puisque, lorsqu'on calcule les valeurs obtenues, on ne fait qu'une projection orthogonale, il nous suffit de projeter l'ensemble des points de l'intervalle pour trouver les nouvelles bornes de l'intervalle. On obtient alors les bornes suivantes pour la l^e composante principale du i^e objet :

$$\underline{y_{il}} = \sum_{j=1}^{p} \min_{\underline{x_{ij}} \leq x_{ij}^r \leq \overline{x_{ij}}} (x_{ij}^r - \overline{x_j^c})\, v_{jl}$$

$$\overline{y_{il}} = \sum_{j=1}^{p} \max_{\underline{x_{ij}} \leq x_{ij}^r \leq \overline{x_{ij}}} (x_{ij}^r - \overline{x_j^c})\, v_{jl}.$$

Nous pouvons remarquer que, pour le calcul de $\underline{y_{il}}$:

- si $v_{jl} < 0$: le minimum est atteint en $\overline{x_{ij}}$
- si $v_{jl} > 0$: le minimum est atteint en $\underline{x_{ij}}$

et pour le calcul de $\overline{y_{il}}$:

- si $v_{jl} < 0$: le maximum est atteint en $\underline{x_{ij}}$
- si $v_{jl} > 0$: le maximum est atteint en $\overline{x_{ij}}$.

Les bornes de la l$^{\text{e}}$ composante principale intervalle Y_l^I pour l'objet i peuvent alors aussi s'écrire :

$$\underline{y_{il}} = \sum_{\{j|v_{jl}<0\}} (\overline{x_{ij}} - \overline{x_j^c})\, v_{jl} + \sum_{\{j|v_{jl}>0\}} (\underline{x_{ij}} - \overline{x_j^c})\, v_{jl}$$

$$\overline{y_{il}} = \sum_{\{j|v_{jl}<0\}} (\underline{x_{ij}} - \overline{x_j^c})\, v_{jl} + \sum_{\{j|v_{jl}>0\}} (\overline{x_{ij}} - \overline{x_j^c})\, v_{jl}.$$

2.5.5 Interprétation des résultats

Nous pouvons également adapter les paramètres d'interprétation dans le cas de la méthode des centres.

$*$ Pour mesurer la qualité de représentation de l'individu i par rapport à la l$^{\text{e}}$ composante principale :

$$COR_I^1(i, Y_l^I) = COR_I^2(i, Y_l^I) = \frac{\omega_{C_i}\, y_{C_i l}^2}{\omega_{C_i}\, d^2(C_i, G)} = \frac{y_{C_i l}^2}{d^2(C_i, G)}$$

où ω_{C_i} est le poids du centre c_i et est donc égal à p_i puisque toute l'information concernant R_i est ramenée sur c_i.

Ce paramètre indique le rapport entre la contribution du centre c_i du i$^{\text{e}}$ hyperrectangle à l'inertie de la l$^{\text{e}}$ composante principale et la contribution de ce centre à l'inertie totale.

$*$ Pour mesurer la contribution de x_i à la variance λ_l de la l$^{\text{e}}$ composante principale :

$$CTR_I(i, Y_l^I) = \frac{\omega_{C_i}}{\lambda_l}\, y_{C_i l}^2$$

* Pour mesurer la contribution de x_i au SST des n centres représentant les n hyperrectangles :

$$INR_I(i) = \frac{\omega_{C_i}\, d^2(C_i, G)}{I_T} = \omega_{C_i}\, \frac{d^2(C_i, G)}{\sum_{j=1}^{p} \lambda_j}.$$

2.5.6 Lien avec la méthode classique

Montrons que l'analyse en composantes principales classique est aussi un cas particulier de la méthode des centres.

Puisque les données classiques sont un cas particulier des données intervalles où la borne inférieure et supérieure de l'intervalle coïncident, on a :

$$x^c_{ij} = x_{ij}.$$

La matrice $\widetilde{X}$ dont les lignes sont les coordonnées des centres est donc la matrice X des données classiques.

On construit alors les composantes principales intervalles :

- $\overline{x^c_j}$ la moyenne de la j^e colonne de $\widetilde{X}$ est égale à $\overline{x_j}$ la moyenne des éléments de la j^e colonne de X puisque $\widetilde{X} = X$.

- La valeur de la l^e composante principale pour le centre c_i est égale à la valeur de la l^e composante principale pour l'objet i.

- On a alors que

  $$\underline{y_{il}} = \overline{y_{il}} = y_{il} = y^c_{il}$$

  puisque le minimum et le maximum ne sont pris que sur la valeur x_{ij}.

De nouveau, cette étape ne modifie pas les valeurs obtenues par la méthode classique et on a bien que l'analyse en composantes principales classique est un cas particulier de la méthode des centres.

2.6 Comparaison des 2 méthodes

Une première remarque à faire concerne les intervalles de petite longueur. Nous avons dans ce cas que les résultats fournis par la méthode des sommets et par la méthode des centres fournissent des résultats très similaires puisque les sommets et les centres ont des coordonnées très semblables.

2.6.1 Analyse inter-classes et intra-classes

Nous allons maintenant comparer les expressions des moyennes, des variances et des covariances que l'on obtient avec les différentes méthodes pour montrer que la méthode des sommets correspond à une analyse inter-objets et intra-objets alors que la méthode des centres correspond uniquement à une analyse inter-objets.

Moyenne des variables Y_j :

Pour la méthode des sommets, la moyenne de la j^e variable que nous noterons Y_j^s correspond à la moyenne de la j^e colonne de la matrice M. Comme nous l'avons vu, cette colonne contient, pour chaque objet i, 2^{q_i-1} fois les bornes de x_{ij}. Nous avons donc :

$$Y_j^s = \sum_{i=1}^{n} \frac{2^{q_i-1}\,(\underline{x}_{ij} + \overline{x}_{ij})}{n.2^{q_i}} = \sum_{i=1}^{n} \frac{\underline{x}_{ij} + \overline{x}_{ij}}{2n}.$$

Pour la méthode des centres, nous avons directement :

$$Y_j^c = \sum_{i=1}^{n} \frac{x_{ij}^c}{n} = \sum_{i=1}^{n} \frac{\underline{x}_{ij} + \overline{x}_{ij}}{2n}.$$

Nous avons donc :

$$Y_j^s = Y_j^c.$$

Les variances :

Calculons la variance v_{jj}^s pour la méthode des sommets :

$$v_{jj}^s = \sum_{i=1}^{n} \underbrace{\sum_{k=1}^{2^{q_i}}}_{\text{sommets de } i} \omega_k \underbrace{(M_i)_{kj}}_{=\underline{x}_{ij} \text{ ou } \overline{x}_{ij}}$$

$$= \sum_{i=1}^{n} \Big(\sum_{k \text{ tels que } (M_i)_{kj}=\underline{x}_{ij}} \frac{1}{n2^{q_i}} \underline{x}_{ij} + \underbrace{\sum_{k \text{ tels que } (M_i)_{kj}=\overline{x}_{ij}} \frac{1}{n2^{q_i}}}_{=(\frac{1}{n2^{q_i}})2^{q_i-1}=\frac{1}{2n}} \overline{x}_{ij} \Big)$$

$$= \underbrace{\sum_{i=1}^{n}}_{\frac{1}{2n}.n=\frac{1}{2}} \Big(\frac{1}{2n} \underline{x}_{ij} + \frac{1}{2n} \underline{x}_{ij} \Big)$$

$$= \frac{1}{2}(\underline{x}_{ij} + \overline{x}_{ij}).$$

La variance v_{jj}^c pour la méthode des centres se trouve directement par :

$$v_{jj}^c = \sum_{i=1}^{n} \frac{1}{4}(\underline{x}_{ij} + \overline{x}_{ij})^2$$

Calculons la différence entre ces deux variances :

$$\begin{aligned} v_{jj}^s - v_{jj}^c &= \Big(\frac{1}{2}(\underline{x}_{ij} + \overline{x}_{ij})\Big) - \Big(\sum_{i=1}^{n} \frac{1}{4}(\underline{x}_{ij} + \overline{x}_{ij})^2\Big) \\ &= \sum_{i=1}^{n} \frac{1}{4}\underline{x}_{ij}^2 + \frac{1}{4}\overline{x}_{ij}^2 - \frac{1}{2}\underline{x}_{ij}\overline{x}_{ij} \\ &= \frac{1}{4}\,(\underline{x}_{ij} + \overline{x}_{ij})^2 \end{aligned}$$

Nous avons donc :

$$v_{jj}^s = v_{jj}^c + \frac{1}{4}\left(\underline{x}_{ij} + \overline{x}_{ij}\right)^2$$
$$v_{jj}^s = v_{jj}^c + e_{jj}$$

e_{jj} sera appelé le facteur d'amplitude de la variable Y_j. Il exprime l'information de variation ou d'imprécision de la variable de type intervalle. Nous noterons E la matrice diagonale regroupant les facteurs d'amplitude des différentes variables Y_j :

$$E = \begin{pmatrix} e_{11} & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & e_{pp} \end{pmatrix}.$$

Les covariances :

Par un calcul similaire au précédent, nous pouvons montrer que les covariances sont identiques dans le cas de la méthode des sommets et dans le cas de la méthode des centres.

Conclusion :

Nous pouvons établir les relations suivantes entre les matrices variance-covariance V^s et V^c :

$$V^s = V^c + E.$$

La méthode des sommets apparaît donc comme une analyse inter et intra-objet puisqu'elle prend en compte :

- la variabilité *entre* les centres des hyperrectangles à travers V^c
- la variabilité *à l'intérieur* des hyperrectangles à travers E.

2.6.2 La complexité des calculs

Dans la méthode des sommets, nous devons appliquer l'analyse en composantes principales classique à la matrice M qui est de dimension $n2^{q_i} \times p$. Lorsque nous travaillons sur un nombre important de variables, le nombre de calculs à effectuer devient assez important.

Dans la méthode des centres, nous appliquons l'analyse en composantes principales classique à la matrice $\widetilde{X}$ de dimension $n \times p$. Le nombre de calculs à effectuer est donc nettement moins important.

2.7 Exemples

2.7.1 Sur des données artificielles

Prenons 5 individus sur lesquels nous allons mesurer 3 variables de type intervalle. Nous obtenons les mesures suivantes :

	Y_1	Y_2	Y_3
1	[0,1]	[1,2]	[0,1]
2	[-1,0]	[0,0]	[1,3]
3	[-2,-1]	[-1,0]	[-2,0]
4	[1,2]	[2,2]	[2,2]
5	[2,3]	[1,1]	[0,]

Nous aurons alors la matrice de données symboliques suivante :

$$\underline{X} = \begin{pmatrix} [0,1] & [1,2] & [0,1] \\ [-1,0] & [0,0] & [1,3] \\ [-2,-1] & [-1,0] & [-2,0] \\ [1,2] & [2,2] & [2,2] \\ [2,3] & [1,1] & [0,0] \end{pmatrix}$$

La méthode des sommets

Résumons brièvement les étapes de la méthode des sommets :

1. On représente chaque donnée x_i par une matrice M_i.
2. On construit la matrice M.
3. On applique l'analyse en composantes principales classique à la matrice M où y_{kl} est la valeur de la l$^{\text{e}}$ composante principale classique pour le sommet k.
4. On obtient les composantes principales intervalles par :

$$y_{il} = [\underline{y_{il}}, \overline{y_{il}}]$$

où

$$\underline{y_{il}} = \min_{k \in L_i} y_{kl}$$

$$\overline{y_{il}} = \max_{k \in L_i} y_{kl}.$$

Première étape : Représentation des données

On représente chaque donnée x_i par une matrice M_i représentant les différents sommets de l'hyperrectangle R_i :

$$M_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 2 & 0 \\ 0 & 2 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 2 & 1 \end{pmatrix} \qquad M_2 = \begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 3 \\ 0 & 0 & 1 \\ 0 & 0 & 3 \end{pmatrix}$$

$$M_3 = \begin{pmatrix} -2 & -1 & -2 \\ -2 & -1 & 0 \\ -2 & 0 & -2 \\ -2 & 0 & 0 \\ -1 & -1 & -2 \\ -1 & -1 & 0 \\ -1 & 0 & -2 \\ -1 & 0 & 0 \end{pmatrix} \qquad M_4 = \begin{pmatrix} 1 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix}$$

$$M_5 = \begin{pmatrix} 2 & 1 & 0 \\ 3 & 1 & 0 \end{pmatrix}.$$

Deuxième étape : Construction de la matrice M

On obtient alors la matrice M suivante :

$$M = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 2 & 0 \\ 0 & 2 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 2 & 1 \\ \\ -1 & 0 & 1 \\ -1 & 0 & 3 \\ 0 & 0 & 1 \\ 0 & 0 & 3 \\ \\ -2 & -1 & -2 \\ -2 & -1 & 0 \\ -2 & 0 & -2 \\ -2 & 0 & 0 \\ & \vdots & \end{pmatrix}$$

$$\begin{pmatrix} \vdots & & \\ & & \\ -1 & -1 & -2 \\ -1 & -1 & 0 \\ -1 & 0 & -2 \\ -1 & 0 & 0 \\ & & \\ 1 & 2 & 2 \\ 2 & 2 & 2 \\ & & \\ 2 & 1 & 0 \\ 3 & 1 & 0 \end{pmatrix}$$

Troisième étape : Application de l'analyse en composantes principales classique

Nous devons alors appliquer l'analyse en composantes principales à la matrice M. On commence donc par calculer le centroïde des points $\overline{m} = (\overline{m_1}, \overline{m_2}, \overline{m_3})$:

$$\overline{m_1} = \frac{1}{24}\sum_{i=1}^{24} x_{i1} = -0.0833$$

$$\overline{m_2} = \frac{1}{24}\sum_{i=1}^{24} x_{i2} = 0.5833$$

$$\overline{m_3} = \frac{1}{24}\sum_{i=1}^{24} x_{i3} = 0.3333.$$

On calcule alors la matrice $\widetilde{M}$:

$$\begin{pmatrix} 0.0833 & 0.4167 & -0.3333 \\ 0.0833 & 0.4167 & 0.6667 \\ 0.0833 & 1.4167 & -0.3333 \\ 0.0833 & 1.4167 & 0.6667 \\ & \vdots & \end{pmatrix}$$

$$\begin{pmatrix} & \vdots & \\ 1.0833 & 0.4167 & -0.3333 \\ 1.0833 & 0.4167 & 0.6667 \\ 1.0833 & 1.4167 & -0.3333 \\ 1.0833 & 1.4167 & 0.6667 \\ \\ -0.9167 & -0.5833 & 0.6667 \\ -0.9167 & -0.5833 & 2.6667 \\ 0.0833 & -0.5833 & 0.6667 \\ 0.0833 & -0.5833 & 2.6667 \\ \\ -1.9167 & -1.5833 & -2.3333 \\ -1.9167 & -1.5833 & -0.3333 \\ -1.9167 & -0.5833 & -2.3333 \\ -1.9167 & -0.5833 & -0.3333 \\ -0.9167 & -1.5833 & -2.3333 \\ -0.9167 & -1.5833 & -0.3333 \\ -0.9167 & -0.5833 & -2.3333 \\ -0.9167 & -0.5833 & -0.3333 \\ \\ 1.0833 & 1.4167 & 1.6667 \\ 2.0833 & 1.4167 & 1.6667 \\ \\ 2.0833 & 0.4167 & -0.3333 \\ 3.0833 & 0.4167 & -0.3333 \end{pmatrix}$$

Et la matrice de covariance S :

$$S = \widetilde{X}' \, \widetilde{X} = \begin{pmatrix} 1.9058 & 1.0507 & 0.7246 \\ 1.0507 & 1.1232 & 0.5797 \\ 0.7246 & 0.5797 & 1.9710 \end{pmatrix}$$

et les valeurs propres et les vecteurs propres de S :

▷ Les valeurs propres :

$$\lambda_1 = 3.2875$$
$$\lambda_2 = 1.3225$$
$$\lambda_3 = 0.3900.$$

▷ Les vecteurs propres :

$$v_1 = \begin{pmatrix} -0.6644 \\ -0.4767 \\ -0.5756 \end{pmatrix}; \qquad v_2 = \begin{pmatrix} 0.5039 \\ 0.2830 \\ -0.8161 \end{pmatrix}; \qquad v_3 = \begin{pmatrix} 0.5519 \\ -0.8322 \\ 0.0522 \end{pmatrix}.$$

Pour évaluer le nombre de composantes principales nécessaires, calculons la variance totale des données ainsi que le pourcentage de variance expliqué par chaque valeur propre :

$$I = 5$$
$$t_1 = 0.6575$$
$$t_2 = 0.2445$$
$$t_3 = 0.0780.$$

On voit alors que deux composantes principales suffisent à expliquer la proportion :

$$T = 0.922$$

de la variance totale des données.

On construit alors nos deux composantes principales :

$$Y_1^* = v_1' \, Y = -0.6644 \, Y_1 - 0.4767 \, Y_2 - 0.5756 \, Y_3$$
$$Y_2^* = v_2' \, Y = 0.5039 \, Y_1 + 0.283 \, Y_2 - 0.8161 \, Y_3.$$

On trouve alors les coordonnées des sommets dans l'espace à 2 dimensions :

$$z_i^* = \begin{pmatrix} -0.6644 & -0.4767 & -0.5756 \\ 0.5039 & 0.283 & -0.8161 \end{pmatrix} (M^*)_i$$

où $(M^*)_i$ est la i$^{\text{e}}$ ligne de M^*.

On obtient les résultats suivants où $R_i(j)$ est le j^e sommet de R_i :

	Y_1^*	Y_2^*
$R_1(1)$	-0.0621	0.4319
$R_1(2)$	-0.6377	-0.3841
$R_1(3)$	-0.5388	0.715
$R_1(4)$	-1.1145	-0.1011
$R_1(5)$	-0.7265	0.9359
$R_1(6)$	-1.3021	0.1198
$R_1(7)$	-1.2032	1.2189
$R_1(8)$	-1.7788	0.4029
$R_2(1)$	0.5034	-1.1711
$R_2(2)$	-0.6479	-2.8032
$R_2(3)$	-0.161	-0.6672
$R_2(4)$	-1.3122	-2.2993
$R_3(1)$	3.3713	0.4901
$R_3(2)$	2.2201	-1.142
$R_3(3)$	2.8946	0.7732
$R_3(4)$	1.7434	-0.8589
$R_3(5)$	2.7069	0.9941
$R_3(6)$	1.5557	-0.6381
$R_3(7)$	2.2302	1.2771
$R_3(8)$	1.079	-0.355
$R_4(1)$	-2.3545	-0.4132
$R_4(2)$	-3.0188	0.0907
$R_5(1)$	-1.3909	1.4398
$R_5(2)$	-2.0553	1.9437

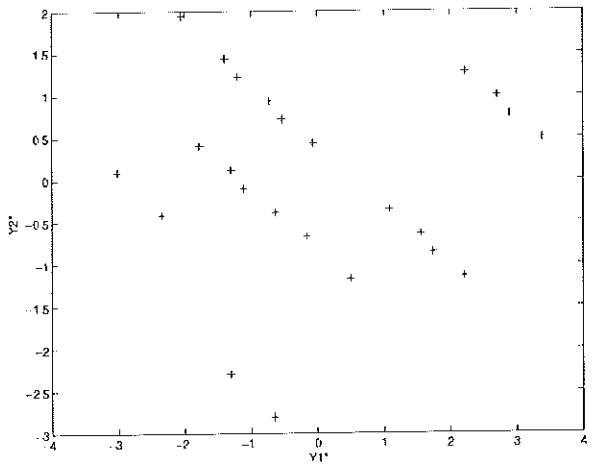


FIG. 2.4 – Représentation des sommets des différents hyperrectangles

Quatrième étape : Construction des composantes principales intervalles

On commence par définir les ensembles L_i :

$$L_1 = \{1, \ldots, 8\}$$
$$L_2 = \{9, \ldots, 12\}$$
$$L_3 = \{13, \ldots, 20\}$$
$$L_4 = \{21, 22\}$$
$$L_5 = \{23, 24\}.$$

On définit ensuite les matrices M_i^* qui nous permettront de trouver les valeurs des composantes principales intervalles pour chaque objet. En effet, nous savons que les bornes de l'intervalle de la j^e composante principale pour l'individu i sont données par la valeur minimale et maximale de la j^e colonne de la matrice M_i.

On a alors :

$$M_1^* = \begin{pmatrix} -0.0621 & 0.4319 \\ -0.6377 & -0.3841 \\ -0.5388 & 0.715 \\ -1.1145 & -0.1011 \\ -0.7265 & 0.9359 \\ -1.3021 & 0.1198 \\ -1.2032 & 1.2189 \\ 1.7788 & 0.4029 \end{pmatrix}$$

$\triangleright\ Y_1^I(1) = [-1.7788, -0.0621] \qquad Y_2^I(1) = [-0.3841, 1.2189]$

$$M_2^* = \begin{pmatrix} 0.5034 & -1.1711 \\ -0.6479 & -2.8032 \\ -0.1610 & -0.6672 \\ -1.3122 & -2.2993 \end{pmatrix}$$

$\triangleright\ Y_1^I(2) = [-1.3122, 0.5034] \qquad Y_2^I(2) = [-2.8032, -0.6672]$

$$M_3^* = \begin{pmatrix} 3.3713 & 0.4901 \\ 2.2201 & -1.142 \\ 2.8946 & 0.7732 \\ 1.7434 & -0.8589 \\ 2.7069 & 0.9941 \\ 1.5557 & -0.6381 \\ 2.2302 & 1.2771 \\ 1.079 & -0.355 \end{pmatrix}$$

$\triangleright\ Y_1^I(3) = [1.07903.3713] \qquad Y_2^I(3) = [-1.1420, 1.2771]$

$$M_4^* = \begin{pmatrix} -2.3545 & -0.4132 \\ -3.0188 & 0.0907 \end{pmatrix}$$

$\triangleright\ Y_1^I(4) = [-3.0188, -2.3545] \qquad Y_2^I(4) = [-0.4132, 0.0907]$

$$M_5^* = \begin{pmatrix} -1.3909 & 1.4398 \\ -2.0553 & 1.9437 \end{pmatrix}$$

$$\triangleright\ Y_1^I(5) = [-2.0553, -1.3909] \qquad Y_2^I(5) = [1.4398, 1.9437]$$

Graphiquement, on a alors :

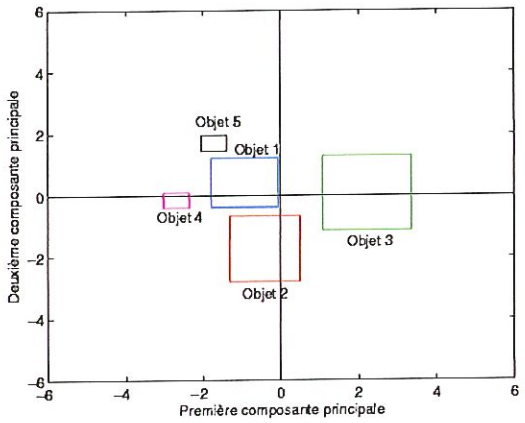


FIG. 2.5 – Représentation des individus dans l'espace formé par les deux composantes principales intervalles

La méthode des centres

Résumons les différentes étapes de la méthode des centres :

1. On calcule les centres $c_i \ \ (i = 1, \ldots, n)$ et on construit la matrice $\widetilde{X}$.
2. On applique l'analyse en composantes principales classique à la matrice $\widetilde{X}$.
3. On construit alors les composantes principales intervalles $y_{il} = [\underline{y}_{il}, \overline{y}_{il}]$ par :

$$\underline{y_{il}} = \sum_{\{j|v_{jl}<0\}} (\overline{x_{ij}} - \overline{x_j^c})\, v_{jl} + \sum_{\{j|v_{jl}>0\}} (\underline{x_{ij}} - \overline{x_j^c})\, v_{jl}$$

$$\overline{y_{il}} = \sum_{\{j|v_{jl}<0\}} (\underline{x_{ij}} - \overline{x_j^c})\, v_{jl} + \sum_{\{j|v_{jl}>0\}} (\overline{x_{ij}} - \overline{x_j^c})\, v_{jl}.$$

Première étape : Construction de $\widetilde{X}$

On reprend nos données de départ :

$$\underline{X} = \begin{pmatrix} [0,1] & [1,2] & [0,1] \\ [-1,0] & [0,0] & [1,3] \\ [-2,-1] & [-1,0] & [-2,0] \\ [1,2] & [2,2] & [2,2] \\ [2,3] & [1,1] & [0,0] \end{pmatrix}.$$

On calcule les centres $c_i = (x_{i1}^c, \ldots, x_{ip}^c)$ de chaque hyperrectangle où :

$$x_{ij}^c = \frac{\underline{x_{ij}} + \overline{x_{ij}}}{2} \ ; \quad j = 1, \ldots, p \ ; \quad i = 1, \ldots, n.$$

On obtient alors :

$$\widetilde{X} = \begin{pmatrix} 0.5 & 1.5 & 0.5 \\ -0.5 & 0 & 2 \\ -1.5 & -0.5 & -1 \\ 1.5 & 2 & 2 \\ 2.5 & 1 & 0 \end{pmatrix}.$$

Deuxième étape : Application de l'analyse en composantes principales classique

On applique alors l'analyse en composantes principales classique à la matrice $\widetilde{X}$. Calculons le centroïde des centres des hyperrectangles $\overline{\widetilde{x}} = (\overline{\widetilde{x_1}}, \overline{\widetilde{x_2}}, \overline{\widetilde{x_3}})$:

$$\overline{\widetilde{x_1}} = \frac{1}{5}\sum_{i=1}^{5}(x_{C_i})_1 = 0.5$$

$$\overline{\widetilde{x_2}} = \frac{1}{5}\sum_{i=1}^{5}(x_{C_i})_2 = 0.8$$

$$\overline{\widetilde{x_3}} = \frac{1}{5}\sum_{i=1}^{5}(x_{C_i})_3 = 0.7.$$

On a alors que :

$$\widetilde{X}^* = \widetilde{X} - \overline{\widetilde{X}} = \begin{pmatrix} 0 & 0.7 & -0.2 \\ -1 & -0.8 & 1.3 \\ -2 & -1.3 & -1.7 \\ 1 & 1.2 & 1.3 \\ 2. & 0.2 & -0.7 \end{pmatrix}.$$

On peut alors calculer la matrice de covariance S :

$$S = \begin{pmatrix} 2.5 & 1.25 & 0.5 \\ 1.25 & 1.075 & 0.6125 \\ 0.5 & 0.6125 & 1.7 \end{pmatrix}.$$

Ainsi que ses valeurs et vecteurs propres :

$$\lambda_1 = 3.5297 \qquad \lambda_2 = 1.4597 \qquad \lambda_3 = 0.2856$$

$$v_1 = \begin{pmatrix} -0.7829 \\ -0.4933 \\ -0.3791 \end{pmatrix} \qquad v_2 = \begin{pmatrix} 0.4386 \\ -0.0056 \\ -0.8987 \end{pmatrix} \qquad v_3 = \begin{pmatrix} 0.4412 \\ -0.8699 \\ 0.2207 \end{pmatrix}.$$

Pour évaluer le nombre de composantes principales nécessaire, calculons :

$$I = 5.2750$$

$$t_1 = 0.6691$$

$$t_2 = 0.2767$$

$$t_3 = 0.0541.$$

On voit alors que deux composantes principales nous suffisent pour expliquer la proportion :

$$T = 0.946$$

de la variance totale des données.

On obtient alors les coordonnées suivantes pour les sommets dans l'espace à deux dimensions formé par les composantes principales :

Centres	Y_1^*	Y_2^*
c_1	-0.2695	0.1758
c_2	0.6847	-1.6024
c_3	2.8516	0.6577
c_4	-1.8677	-0.7363
c_5	-1.3991	1.5052

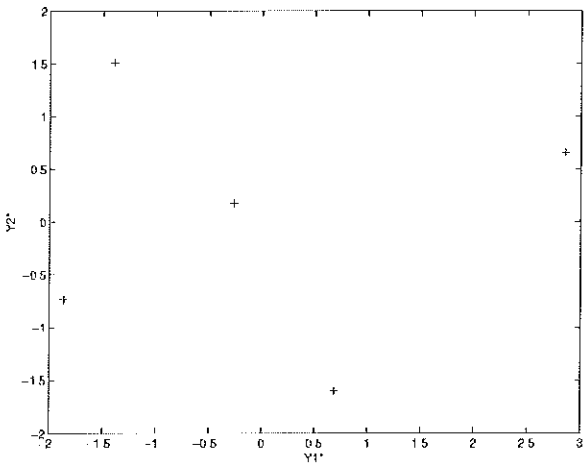


FIG. 2.6 – Représentation des centres des individus

Troisième étape : Construction des composantes principales intervalles

Pour calculer les valeurs des composantes principales intervalles pour chaque individu i représenté par le centre c_i, il nous suffit alors de calculer :

$$\underline{y_{ij}} = \sum_{\{k|v_{kj}<0\}} (\overline{x_{ik}} - \overline{x_k^c})\, v_{kj} + \sum_{\{k|v_{kj}>0\}} (\underline{x_{ik}} - \overline{x_k^c})\, v_{kj}$$
$$\overline{y_{ij}} = \sum_{\{k|v_{kj}<0\}} (\underline{x_{ik}} - \overline{x_k^c})\, v_{kj} + \sum_{\{k|v_{kj}>0\}} (\overline{x_{ik}} - \overline{x_k^c})\, v_{kj}$$

où :

$$\overline{x}^c = \begin{pmatrix} 0.5 & 0.8 & 0.7 \end{pmatrix}$$

$$v_1' = \begin{pmatrix} 0.7829 & 0.4933 & 0.3791 \end{pmatrix}$$
$$v_2' = \begin{pmatrix} 0.4386 & -0.0056 & -0.8987 \end{pmatrix}.$$

On a ainsi, par exemple, pour le premier individu :

$$\underline{y_{11}} = \sum_{k=1}^{3} (\overline{x_{1k}} - \overline{x_k}^c)\, v_{k1}$$
$$\underline{y_{11}} = -0.55816$$

$$\overline{y_{11}} = \sum_{k=1}^{3} (\underline{x_{1k}} - \overline{x_k}^c)\, v_{k1}$$
$$\overline{y_{11}} = 1.09714$$

$$\underline{y_{12}} = (\underline{x_{11}} - \overline{x_1}^c)\, v_{12} + \sum_{k=2}^{3} (\overline{x_{1k}} - \overline{x_k}^c)\, v_{k2}$$
$$\underline{y_{12}} = -0.49563$$

$$\overline{y_{12}} = (\overline{x_{11}} - \overline{x_1}^c)\, v_{12} + \sum_{k=2}^{3} (\underline{x_{1k}} - \overline{x_k}^c)\, v_{k2}$$
$$\overline{y_{12}} = 0.84727$$

Et ainsi de suite pour tous les individus. On obtient alors :

Individus	Y_1^I	Y_2^I
1	[-1.0971 ; 0.5582]	[-0.4956 ; 0.8473]
2	[-0.0858 ; 1.4553]	[-2.7204 ; -0.4845]
3	[1.8384 ; 3.8687]	[-0.4631 ; 1.7784]
4	[-2.2591 ; -1.4762]	[-0.9556 ; -0.5170]
5	[-1.7906 ; -1.0077]	[1.2859 ; 1.7245]

Et la représentation suivante :

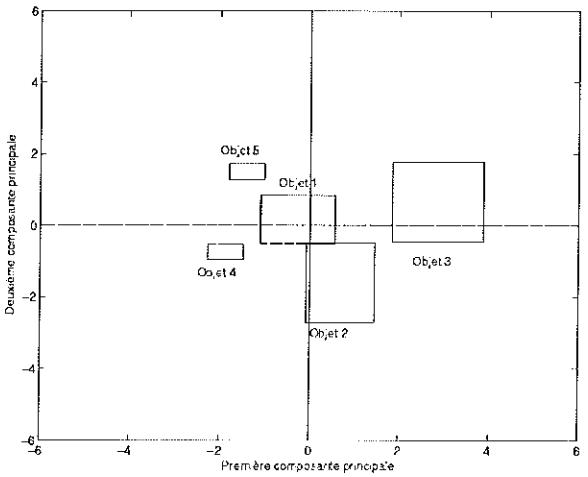


FIG. 2.7 – Représentation des individus dans l'espace formé par les deux premières composantes principales

2.7.2 Sur des données réelles : les données d'Ichino

Présentation des données

Il s'agit de la description de 8 types d'huiles par 4 variables intervalles :

- Y_1 : Gravité spécifique
- Y_2 : Point de congélation
- Y_3 : Valeur iodine
- Y_4 : Valeur de saponification

Nous avons alors les données suivantes :

	Y_1	Y_2	Y_3	Y_4
Linseed	[0.93 , 0.94]	[-27 , -18]	[170 , 204]	[118 , 196]
Perilla	[0.93 , 0.94]	[-5 , -4]	[192 , 208]	[188 , 197]
Cotton	[0.92 , 0.92]	[-6 , -1]	[99 , 113]	[189 , 198]
Sesame	[0.92 , 0.93]	[-6 , -4]	[104 , 116]	[187 , 193]
Camellia	[0.92 , 0.92]	[-21 , -15]	[80 , 82]	[189 , 193]
Olive	[0.91 , 0.92]	[0 ,6]	[79 , 90]	[187 , 196]
Beef	[0.86 , 0.87]	[30 , 38]	[40 , 48]	[190 , 199]
Hog	[0.86 , 0.86]	[22 , 32]	[53 , 77]	[190 , 202]

Notons par :

- 1 : Linseed
- 2 : Perilla
- 3 : Cotton
- 4 : Sesame
- 5 : Camellia
- 6 : Olive
- 7 : Beef
- 8 : Hog

La méthode des sommets

Première étape : Représentation des données

On représente alors chaque huile i par une matrice M_i :

$$M_1 = \begin{pmatrix} 0.93 & -27 & 170 & 118 \\ 0.93 & -27 & 170 & 196 \\ 0.93 & -27 & 204 & 118 \\ 0.93 & -27 & 204 & 196 \\ 0.93 & -18 & 170 & 118 \\ 0.93 & -18 & 170 & 196 \\ 0.93 & -18 & 204 & 118 \\ 0.93 & -18 & 204 & 196 \\ 0.94 & -27 & 170 & 118 \\ 0.94 & -27 & 170 & 196 \\ 0.94 & -27 & 204 & 118 \\ 0.94 & -27 & 204 & 196 \\ 0.94 & -18 & 170 & 118 \\ 0.94 & -18 & 170 & 196 \\ 0.94 & -18 & 204 & 118 \\ 0.94 & -18 & 204 & 196 \end{pmatrix} \quad M_2 = \begin{pmatrix} 0.93 & -5 & 192 & 188 \\ 0.93 & -5 & 192 & 197 \\ 0.93 & -5 & 208 & 188 \\ 0.93 & -5 & 208 & 197 \\ 0.93 & -4 & 192 & 188 \\ 0.93 & -4 & 192 & 197 \\ 0.93 & -4 & 208 & 188 \\ 0.93 & -4 & 208 & 197 \\ 0.94 & -5 & 192 & 188 \\ 0.94 & -5 & 192 & 197 \\ 0.94 & -5 & 208 & 188 \\ 0.94 & -5 & 208 & 197 \\ 0.94 & -4 & 192 & 188 \\ 0.94 & -4 & 192 & 197 \\ 0.94 & -4 & 208 & 188 \\ 0.94 & -4 & 208 & 197 \end{pmatrix}$$

$$M_3 = \begin{pmatrix} 0.92 & -6 & 99 & 189 \\ 0.92 & -6 & 99 & 198 \\ 0.92 & -6 & 113 & 189 \\ 0.92 & -6 & 113 & 198 \\ 0.92 & -1 & 99 & 189 \\ 0.92 & -1 & 99 & 198 \\ 0.92 & -1 & 113 & 189 \\ 0.92 & -1 & 113 & 198 \end{pmatrix} \quad M_4 = \begin{pmatrix} 0.92 & -6 & 104 & 187 \\ 0.92 & -6 & 104 & 193 \\ 0.92 & -6 & 116 & 187 \\ 0.92 & -6 & 116 & 193 \\ 0.92 & -4 & 104 & 187 \\ 0.92 & -4 & 104 & 193 \\ 0.92 & -4 & 116 & 187 \\ 0.92 & -4 & 116 & 193 \\ 0.93 & -6 & 104 & 187 \\ 0.93 & -6 & 104 & 193 \\ 0.93 & -6 & 116 & 187 \\ 0.93 & -6 & 116 & 193 \\ 0.93 & -4 & 104 & 187 \\ 0.93 & -4 & 104 & 193 \\ 0.93 & -4 & 116 & 187 \\ 0.93 & -4 & 116 & 193 \end{pmatrix}$$

$$M_5 = \begin{pmatrix} 0.92 & -21 & 80 & 189 \\ 0.92 & -21 & 80 & 193 \\ 0.92 & -21 & 82 & 189 \\ 0.92 & -21 & 82 & 193 \\ 0.92 & -15 & 80 & 189 \\ 0.92 & -15 & 80 & 193 \\ 0.92 & -15 & 82 & 189 \\ 0.92 & -15 & 82 & 193 \end{pmatrix} \quad M_6 = \begin{pmatrix} 0.91 & 0 & 79 & 187 \\ 0.91 & 0 & 79 & 196 \\ 0.91 & 0 & 90 & 187 \\ 0.91 & 0 & 90 & 196 \\ 0.91 & 6 & 79 & 187 \\ 0.91 & 6 & 79 & 196 \\ 0.91 & 6 & 90 & 187 \\ 0.91 & 6 & 90 & 196 \\ 0.92 & 0 & 79 & 187 \\ 0.92 & 0 & 79 & 196 \\ 0.92 & 0 & 90 & 187 \\ 0.92 & 0 & 90 & 196 \\ 0.92 & 6 & 79 & 187 \\ 0.92 & 6 & 79 & 196 \\ 0.92 & 6 & 90 & 187 \\ 0.92 & 6 & 90 & 196 \end{pmatrix}$$

$$M_7 = \begin{pmatrix} 0.86 & 30 & 40 & 190 \\ 0.86 & 30 & 40 & 199 \\ 0.86 & 30 & 48 & 190 \\ 0.86 & 30 & 48 & 199 \\ 0.86 & 38 & 40 & 190 \\ 0.86 & 38 & 40 & 199 \\ 0.86 & 38 & 48 & 190 \\ 0.86 & 38 & 48 & 199 \\ 0.87 & 30 & 40 & 190 \\ 0.87 & 30 & 40 & 199 \\ 0.87 & 30 & 48 & 190 \\ 0.87 & 30 & 48 & 199 \\ 0.87 & 38 & 40 & 190 \\ 0.87 & 38 & 40 & 199 \\ 0.87 & 38 & 48 & 190 \\ 0.87 & 38 & 48 & 199 \end{pmatrix} \quad M_8 = \begin{pmatrix} 0.86 & 22 & 53 & 190 \\ 0.86 & 22 & 53 & 202 \\ 0.86 & 22 & 77 & 190 \\ 0.86 & 22 & 77 & 202 \\ 0.86 & 32 & 53 & 190 \\ 0.86 & 32 & 53 & 202 \\ 0.86 & 32 & 77 & 190 \\ 0.86 & 32 & 77 & 202 \end{pmatrix}$$

Deuxième étape : Construction de la matrice M

On regroupe ensuite toutes les matrices M_i dans une matrice M qui sera de dimensions 104×4 :

$$M = \begin{pmatrix} M_1 \\ M_2 \\ M_3 \\ M_4 \\ M_5 \\ M_6 \\ M_7 \\ M_8 \end{pmatrix}$$

Troisième étape : Application de l'analyse en composantes principales classique

On applique alors l'analyse en composantes principales classique à M. Pour cela, on calcule la matrice de variance covariance S de dimension 4×4. On obtient :

$$S = 10^3 * \begin{pmatrix} 0 & -0.0005 & 0.0012 & -0.0002 \\ -0.0005 & 0.3628 & -0.7477 & 0.1514 \\ 0.0012 & -0.7477 & 3.2330 & -0.4195 \\ -0.0002 & 0.1514 & -0.4195 & 0.4204 \end{pmatrix}.$$

On calcule ensuite les valeurs propres et les vecteurs propres de S :

$$\begin{aligned} \lambda_1 &= 3480.3 \\ \lambda_2 &= 367.7607 \\ \lambda_3 &= 168.0191 \\ \lambda_4 &= 0.0001 \end{aligned}$$

$$v_1 = \begin{pmatrix} 0.0004 \\ -0.2374 \\ 0.9608 \\ -0.1435 \end{pmatrix}; \quad v_2 = \begin{pmatrix} 0 \\ 0.2012 \\ 0.1931 \\ 0.9603 \end{pmatrix};$$

$$v_3 = \begin{pmatrix} 0.001 \\ -0.9503 \\ -0.1991 \\ 0.2392 \end{pmatrix}; \quad v_4 = \begin{pmatrix} 1 \\ 0.0011 \\ -0.0002 \\ -0.0001 \end{pmatrix}.$$

La variance totale et les pourcentages de variance expliquée par chaque valeur propre sont donnés par :

$$\begin{aligned} I &= 4016.1 \\ t_1 &= 0.8666 \\ t_2 &= 0.0916 \\ t_3 &= 0.0418 \\ t_4 &= 2.6194.10^{-8}. \end{aligned}$$

On décide alors de garder 2 composantes principales qui correspondent à λ_1 et λ_2 :

$$\begin{aligned} Y_1^* &= 0.0004Y_1 - 0.2374Y_2 + 0.9608Y_3 - 0.1435Y_4 \\ Y_2^* &= 0.2012Y_2 + 0.1931Y_3 + 0.9603Y_4. \end{aligned}$$

Nous pouvons alors calculer les nouvelles coordonnées des sommets dans l'espace à 2 dimensions où $R_i(r)$ est le r^e sommet de l'hyperrectangle R_i :

	Y_1^*	Y_2^*
$R_1(1)$	68.8381	-61.4433
$R_1(2)$	57.6487	13.4615
$R_1(3)$	101.5042	-54.8776
$R_1(4)$	90.3147	20.0272
$R_1(5)$	66.7017	-59.6321
$R_1(6)$	55.5123	15.2727
$R_1(7)$	99.3678	-53.0664
$R_1(8)$	88.1783	21.8384
$R_1(9)$	68.8381	-61.4433
$R_1(10)$	57.6487	13.4615

	Y_1^*	Y_2^*
$R_1(11)$	101.5042	-54.8776
$R_1(12)$	90.3148	20.0272
$R_1(13)$	66.7017	-59.6321
$R_1(14)$	55.5123	15.2727
$R_1(15)$	99.3678	-53.0664
$R_1(16)$	88.1783	21.8384
$R_2(1)$	74.7109	14.4547
$R_2(2)$	73.4198	23.0975
$R_2(3)$	90.0831	17.5444
$R_2(4)$	88.7920	26.1873
$R_2(5)$	74.4735	14.6559
$R_2(6)$	73.1824	23.2988
$R_2(7)$	89.8457	17.7457
$R_2(8)$	88.5547	26.3885
$R_2(9)$	74.7109	14.4547
$R_2(10)$	73.4198	23.0975
$R_2(11)$	90.0831	17.5444
$R_2(12)$	88.7920	26.1873
$R_2(13)$	74.4735	14.6559
$R_2(14)$	73.1824	23.2988
$R_2(15)$	89.8457	17.7457
$R_2(16)$	88.5547	26.3885
$R_3(1)$	-14.5465	-2.7453
$R_3(2)$	-15.8376	5.8975
$R_3(3)$	-1.0958	-0.0418
$R_3(4)$	-2.3869	8.6010
$R_3(5)$	-15.7334	-1.7391
$R_3(6)$	-17.0245	6.9037
$R_3(7)$	-2.2827	0.9644
$R_3(8)$	-3.5738	9.6072

	Y_1^*	Y_2^*
$R_4(1)$	-9.4558	-3.7004
$R_4(2)$	-10.3165	2.0615
$R_4(3)$	2.0734	-1.3831
$R_4(4)$	1.2127	4.3788
$R_4(5)$	-9.9305	-3.2980
$R_4(6)$	-10.7912	2.4640
$R_4(7)$	1.5987	-0.9806
$R_4(8)$	0.7379	4.7813
$R_4(9)$	-9.4558	-3.7004
$R_4(10)$	-10.3165	2.0615
$R_4(11)$	2.0734	-1.3831
$R_4(12)$	1.2127	4.3788
$R_4(13)$	-9.9305	-3.2980
$R_4(14)$	-10.7912	2.4640
$R_4(15)$	1.5987	-0.9806
$R_4(16)$	0.7380	4.7813
$R_5(1)$	-29.2404	-9.4331
$R_5(2)$	-29.8142	-5.5918
$R_5(3)$	-27.3189	-9.0469
$R_5(4)$	-27.8927	-5.2056
$R_5(5)$	-30.6647	-8.2256
$R_5(6)$	-31.2385	-4.3843
$R_5(7)$	-27.3189	-9.0469
$R_5(8)$	-29.8142	-5.5918
$R_6(1)$	-34.8992	-7.3207
$R_6(2)$	-36.1903	1.3222
$R_6(3)$	-24.3308	-5.1965
$R_6(4)$	-25.6219	3.4464
$R_6(5)$	-36.3235	-6.1132
$R_6(6)$	-37.6146	2.5296
$R_6(7)$	-25.7550	-3.9890
$R_6(8)$	-27.0461	4.6538

	Y_1^*	Y_2^*
$R_6(9)$	-34.8992	-7.3207
$R_6(10)$	-36.1903	1.3222
$R_6(11)$	-24.3308	-5.1965
$R_6(12)$	-25.6219	3.4464
$R_6(13)$	-36.3235	-6.1132
$R_6(14)$	-37.6146	2.5296
$R_6(15)$	-25.7550	-3.9890
$R_6(16)$	-27.0461	4.6538
$R_7(1)$	-79.9208	-5.9337
$R_7(2)$	-81.2119	2.7092
$R_7(3)$	-72.2347	-4.3888
$R_7(4)$	-73.5258	4.2541
$R_7(5)$	-81.8199	-4.3237
$R_7(6)$	-83.1109	4.3191
$R_7(7)$	-74.1337	-2.7788
$R_7(8)$	-75.4248	5.8640
$R_7(9)$	-79.9208	-5.9337
$R_7(10)$	-81.2119	2.7092
$R_7(11)$	-72.2347	-4.3888
$R_7(12)$	-73.5258	4.2541
$R_7(13)$	-81.8199	-4.3237
$R_7(14)$	-83.1109	4.3191
$R_7(5)$	-74.1337	-2.7788
$R_7(16)$	-75.4248	5.8640
$R_8(1)$	-65.5318	-5.0332
$R_8(2)$	-67.2533	6.4906
$R_8(3)$	-42.4734	-0.3986
$R_8(4)$	-44.1949	11.1252
$R_8(5)$	-67.9056	-3.0208
$R_8(6)$	-69.6271	8.5031
$R_8(7)$	-44.8472	1.6138
$R_8(8)$	-46.5687	13.1377

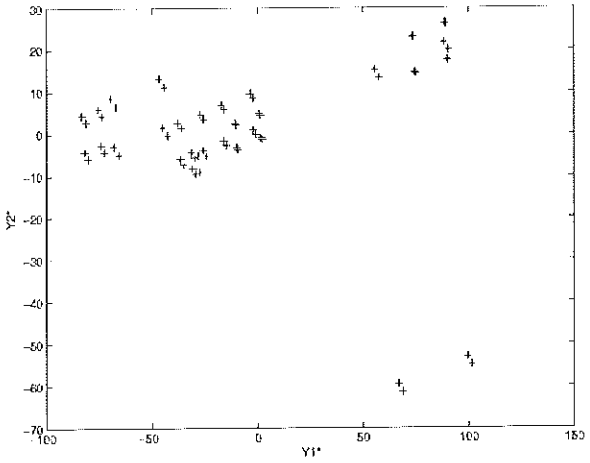


FIG. 2.8 – Représentation des sommets dans l'espace formé par les deux premières composantes principales

Quatrième étape : Construction des composantes principales intervalles

Pour obtenir les composantes principales intervalles pour chaque type d'huile, il suffit alors de prendre le minimum et le maximum des coordonnées des sommets correspondant à ce type d'huile. On a alors :

$$Y_1^I(1) = [55.5123, -101.5042] \qquad Y_2^I(1) = [-61.4433, 21.8384]$$
$$Y_1^I(2) = [73.1824, 90.0831] \qquad Y_2^I(2) = [14.4547, 26.3885]$$
$$Y_1^I(3) = [-17.02459, -1.0958] \qquad Y_2^I(3) = [-2.7453, 8.6010]$$

$$Y_1^I(4) = [-10.7912, 2.0734] \qquad Y_2^I(4) = [-3.7004, 4.3788]$$
$$Y_1^I(5) = [-31.2385, -27.3189] \qquad Y_2^I(5) = [-9.4331, -4.3843]$$
$$Y_1^I(6) = [-37.6146, -24.3308] \qquad Y_2^I(6) = [-7.3207, 4.6538]$$
$$Y_1^I(7) = [-83.1109, -72.2347] \qquad Y_2^I(7) = [-5.9337, 5.9337]$$
$$Y_1^I(8) = [-69.6271, -42.4734] \qquad Y_2^I(8) = [-5.0332, 13.1377].$$

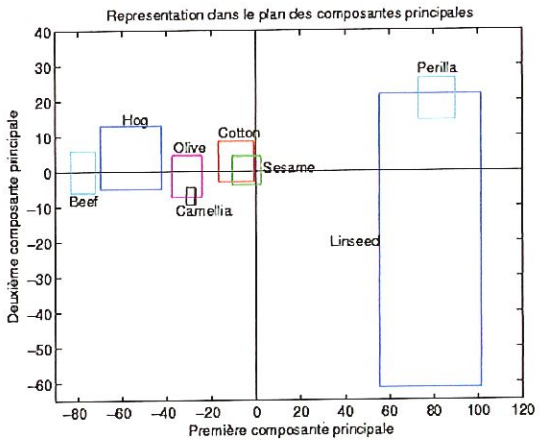


FIG. 2.9 – Représentation des huiles dans l'espace formé par les deux premières composantes principales

La méthode des centres

Première étape : Construction de $\widetilde{X}$

On calcule la matrice des centres $\widetilde{X}$:

$$\widetilde{X} = \begin{pmatrix} 0.935 & -22.5 & 187 & 157 \\ 0.935 & -4.5 & 200 & 192.5 \\ 0.92 & -3.5 & 106 & 193.5 \\ 0.925 & -5 & 110 & 190 \\ 0.92 & -18 & 81 & 191 \\ 0.915 & 3 & 84.5 & 191.5 \\ 0.965 & 34 & 44 & 194.5 \\ 0.86 & 27 & 65 & 196 \end{pmatrix}.$$

Deuxième étape : Application de l'analyse en composantes principales classique

La matrice de variance covariance est donnée par :

$$S = 10^3 * \begin{pmatrix} 0 & -0.0005 & 0.0013 & -0.0002 \\ -0.0005 & 0.3954 & -0.7334 & 0.1484 \\ 0.00013 & -0.7334 & 3.1309 & -0.4241 \\ -0.0002 & 0.1484 & -0.4241 & 0.1632 \end{pmatrix}.$$

Les valeurs propres et les vecteurs propres de S sont :

$$\begin{aligned} \lambda_1 &= 3377.4 \\ \lambda_2 &= 224.3179 \\ \lambda_3 &= 87.8006 \\ \lambda_4 &= 0.00006 \end{aligned}$$

$$v_1 = \begin{pmatrix} 0.0004 \\ -0.2430 \\ 0.9602 \\ -0.1379 \end{pmatrix}; \quad v_2 = \begin{pmatrix} 0.0009 \\ -0.9113 \\ -0.2747 \\ -0.3068 \end{pmatrix}; \quad v_3 = \begin{pmatrix} 0.0009 \\ -0.3325 \\ 0.0511 \\ 0.9417 \end{pmatrix}; \quad v_4 = \begin{pmatrix} 1 \\ 0.0012 \\ -0.0002 \\ -0.0006 \end{pmatrix}.$$

On peut alors calculer la variance totale des données et le pourcentage de variance expliqué par chacune des valeurs propres :

$$\begin{aligned} I &= 3689.6 \\ t_1 &= 0.9154 \\ t_2 &= 0.0608 \\ t_3 &= 0.0238 \\ t_4 &= 1.7867.10^{-8}. \end{aligned}$$

On décide alors de garder 2 composantes principales :

$$Y_1^* = 0.0004Y_1 - 0.243Y_2 + 0.9602Y_3 - 0.1379Y_4$$
$$Y_2^* = 0.0009Y_1 - 0.9113Y_2 - 0.2747Y_3 - 0.3068Y_4.$$

On trouve alors les coordonnées des centres dans l'espace à 2 dimensions :

	Y_1^*	Y_2^*
c_1	84.3294	10.0500
c_2	87.5419	-20.8155
c_3	-3.0952	3.7877
c_4	1.5927	5.1296
c_5	-23.2311	24.6355
c_6	-25.0426	4.384
c_7	-71.8765	-13.6607
c_8	-50.2187	-13.5106

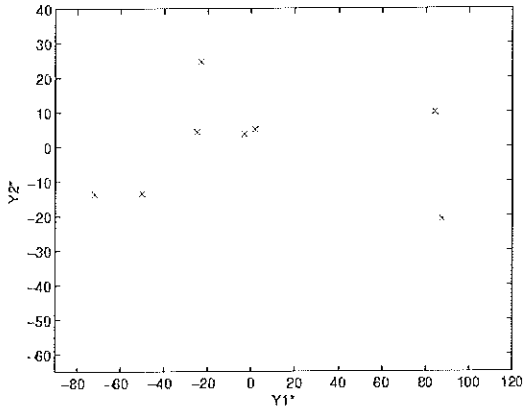


FIG. 2.10 – Représentation des centres dans l'espace formé par les deux premières composantes principales

Troisième étape : Construction des composantes principales intervalles

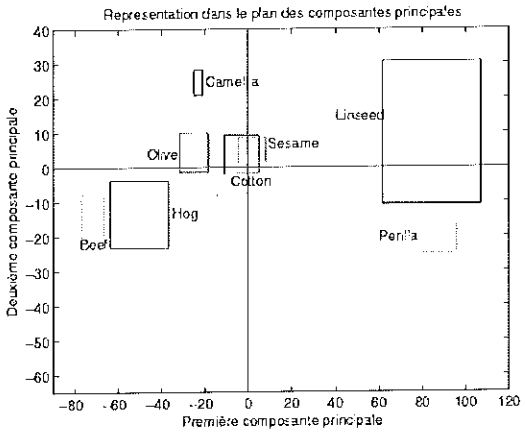
Pour calculer les valeurs des composantes principales intervalles pour chaque individu i représenté par le centre c_i, il nous suffit alors de calculer :

$$\underline{y_{ij}} = \sum_{\{k|v_{kj}<0\}} (\overline{x_{ik}} - \overline{x_k^c})\, v_{kj} + \sum_{\{k|v_{kj}>0\}} (\underline{x_{ik}} - \overline{x_k^c})\, v_{kj}$$

$$\overline{y_{ij}} = \sum_{\{k|v_{kj}<0\}} (\underline{x_{ik}} - \overline{x_k^c})\, v_{kj} + \sum_{\{k|v_{kj}>0\}} (\overline{x_{ik}} - \overline{x_k^c})\, v_{kj}$$

Et on obtient alors :

	Y_1^I	Y_2^I
Linseed	[61.5346 , 107.1242]	[-10.6860 , 30.7861]
Perilla	[79.1184 , 95.9654]	[-24.8494 , -16.7817]
Cotton	[-11.0445 , 4.8541]	[-1.7940 , 9.3694]
Sesame	[-4.8251 , 8.0105]	[1.6498 , 8.6095]
Camellia	[-25.1961 , -21.2661]	[21.0133 , 28.2576]
Olive	[-31.6731 , -18.4121]	[-1.2413 , 10.0093]
Beef	[-77.3097 , -66.4432]	[-19.7852 , -7.5362]
Hog	[-63.7832 , -36.6541]	[-23.2041 , -3.8171]



Troisième étape : Construction des composantes principales intervalles

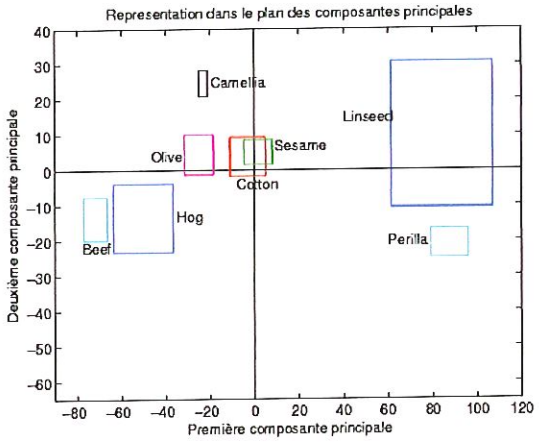
Pour calculer les valeurs des composantes principales intervalles pour chaque individu i représenté par le centre c_i, il nous suffit alors de calculer :

$$\underline{y_{ij}} = \sum_{\{k|v_{kj}<0\}} (\overline{x_{ik}} - \overline{x_k^c})\, v_{kj} + \sum_{\{k|v_{kj}>0\}} (\underline{x_{ik}} - \overline{x_k^c})\, v_{kj}$$

$$\overline{y_{ij}} = \sum_{\{k|v_{kj}<0\}} (\underline{x_{ik}} - \overline{x_k^c})\, v_{kj} + \sum_{\{k|v_{kj}>0\}} (\overline{x_{ik}} - \overline{x_k^c})\, v_{kj}$$

Et on obtient alors :

	Y_1^I	Y_2^I
Linseed	[61.5346 , 107.1242]	[-10.6860 , 30.7861]
Perilla	[79.1184 , 95.9654]	[-24.8494 , -16.7817]
Cotton	[-11.0445 , 4.8541]	[-1.7940 , 9.3694]
Sesame	[-4.8251 , 8.0105]	[1.6498 , 8.6095]
Camellia	[-25.1961 , -21.2661]	[21.0133 , 28.2576]
Olive	[-31.6731 , -18.4121]	[-1.2413 , 10.0093]
Beef	[-77.3097 , -66.4432]	[-19.7852 , -7.5362]
Hog	[-63.7832 , -36.6541]	[-23.2041 , -3.8171]



2.7.3 Interprétation des résultats

Calculons les corrélations entre les quatre variables initiales $Y_1, \ldots, Y_4$ et les deux composantes principales Y_1^I et Y_2^I pour la méthode des sommets et la méthode des centres :

	Y_1^I	Y_2^I
Y_1	-0.8024	-0.122
Y_2	0.7324	0.2906
Y_3	-0.9976	-0.0413
Y_4	0.4021	0.9341

	Y_1^I	Y_2^I
Y_1	-0.7979	-0.4493
Y_2	0.7102	0.6864
Y_3	-0.9973	0.0735
Y_4	0.6273	0.3597

La première composante principale est essentiellement corrélée avec Y_3 (valeur iodine) tandis que la deuxième composante principale l'est avec Y_4 (indice de saponification).

Les huiles à valeur iodine importante auront donc une valeur négative pour Y_1^I et les huiles avec un indice de saponification important auront une valeur plus importante pour la deuxième composante principale.

Nous pouvons remarquer que les huiles Perilla et Linseed sont regroupées. Elles ont en effet une valeur iodine plus faible que toutes les autres. Linseed a en plus un grand intervalle sur Y_2^I puisque sa valeur de saponification est fort variable.

Les huiles Cotton et Sésame sont elles aussi fort regroupées puisqu'elles ont des valeurs assez similaires pour les quatre variables.

Comparaison des résultats obtenus avec les deux méthodes

Les deux méthodes nous donnent des résultats relativement similaires : nous retrouvons à chaque fois les groupes d'huiles Perilla-Linseed, Cotton-Sesame ainsi que celui Beef-Hog.

La différence essentielle vient de l'huile Camellia au niveau de la deuxième composante principale et s'explique par les corrélations. En effet, dans la méthode des sommets, la deuxième composante principale est surtout corrélée avec la variable Y_4 alors que dans la méthode des sommets, elle l'est principalement avec la variables Y_2. Or, si nous examinons les valeurs des variables pour Camellia, nous pouvons remarquer qu'elle se distingue de l'huile d'Olive par son point de congélation c'est-à-dire par sa valeur de Y_2. Elle a donc une valeur plus importante pour sa deuxième composante principale dans le cas

de la méthode des centres.

En ce qui concerne le pourcentage de variance expliquée, nous pouvons voir que la méthode des centres explique un pourcentage plus élevé que la méthode des sommets.

Deuxième partie

L'analyse factorielle discriminante

Deuxième partie

L'analyse factorielle discriminante

Chapitre 3

L'analyse factorielle discriminante classique

3.1 Introduction

Dans cette partie, nous allons voir une deuxième méthode qui nous permettra de diminuer le nombre de variables avec lesquelles nous devons travailler : l'analyse factorielle discriminante.

Nous allons également chercher des axes que nous appellerons axes factoriels qui détermineront un plan factoriel sur lequel nous projetterons nos données initiales.

Ici, nous envisagerons en plus l'aspect classificatoire : les objets appartiendront à différentes classes. Les axes factoriels seront donc construits de manière à maximiser les variances inter-classes mais aussi de manière à minimiser la variance intra-classes pour pouvoir retrouver les différentes classes sur le plan factoriel.

Dans ce chapitre, nous commencerons par quelques définitions pour en déduire la méthode qui nous permettra de construire les axes factoriels. Il nous faudra ensuite définir une règle d'affectation afin de reconstituer les différentes classes.

Pour évaluer la qualité des résultats que nous avons obtenus, nous définirons les notions de taux de bons et de mauvais classements.

Nous terminerons également ce chapitre par un exemple sur des données artificielles pour illustrer les différentes étapes de la méthode et par un exemple sur des données réelles dont nous pourrons interpréter les résultats.

3.2 Les données

Considérons :

- n objets appartenant à un ensemble $E = \{1, \ldots, n\}$ appelé ensemble d'apprentissage
- $Y_1, \ldots, Y_p$, p variables explicatives *quantitatives* dont les domaines $\mathcal{Y}_j$ sont $\mathbb{R}$
- n données $x_1, \ldots, x_n \in \mathbb{R}^p$
- m classes $\Pi_1, \ldots, \Pi_m$ auxquelles appartiennent les différents objets.

Chaque individu de E appartient à une des m classes et leurs appartenances sont connues *a priori*. Elles sont décrites par une variable *nominale* c sur E avec m catégories :

$$c(k) = t \text{ si } k \in \Pi_t, \quad t = 1, \ldots, m.$$

Comme pour l'analyse en composantes principales classique, on commence avec une matrice de données classiques $X = (x_{kj})_{n \times p}$ avec $x_{kj} = Y_j(k)$

$$X = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1' \\ \vdots \\ x_k' \\ \vdots \\ x_n' \end{pmatrix}.$$

Nous supposerons, sans perdre de généralité, que les variables sont centrées c'est-à-dire que :

$$\sum_{i=1}^{n} x_{ij} = 0 \; ; \quad j = 1, \ldots, p$$

Sinon, nous pouvons définir une matrice $\widetilde{X}$ comme pour l'analyse en composantes principales.

On définit également une matrice $C = (c_{kt})$ de dimension $n \times m$ associée à la variable c telle que :

$$c_{kt} = \begin{cases} 1 & \text{si } c(k) = t \text{ c'est-à-dire si } k \in \Pi_t \\ 0 & \text{sinon.} \end{cases}$$

Cette matrice indique donc l'appartenance de chaque objet aux différents groupes. Chaque ligne représente un individu, elle contient donc un seul 1.

Nous noterons par C_t l'ensemble des éléments de E qui appartiennent à la classe Π_t et par $n_t = |C_t|$ le nombre de ces éléments.

Nous considérons ici que chaque individu k a le même poids $p_k = \frac{1}{n}$. Chaque classe a alors un poids égal à $\frac{n_t}{n}$.

3.3 Quelques définitions

Si nous voulons retrouver au mieux les différentes classes sur le plan factoriel, les variables discriminantes doivent être construites de manière à maximiser la variance inter-classes et à minimiser la variance intra-classes.

Pour en arriver à la définition des variances inter et intra-classes, il nous faut encore voir quelques définitions.

3.3.1 Matrice de covariance et matrice de poids

Puisque X est une matrice centrée, la matrice de covariance de X est donnée par :

$$\boxed{T = \left(\frac{1}{n} \sum_{i=1}^{n} x_{ij}\, x_{il}\right)_{p \times p} = X'HX}$$

où H est la matrice des poids des n individus. Nous avons supposé que chaque élément a le même poids donc :

$$H = \frac{1}{n}\, I_n.$$

La matrice diagonale des poids $\frac{n_t}{n}$ des classes sera notée Q et sera définie par :

$$\boxed{(Q)_{n\times m} = \mathrm{diag}\left(\tfrac{n_1}{n}, \ldots, \tfrac{n_m}{n}\right) = C'HC}$$

Preuve :

$$C'HC = C'\left(\frac{1}{n} I_n\right)C = \frac{1}{n}\, C'C$$

Or, l'élément (t, t') de $C'C$ est :

$$(C'C)_{tt'} = \sum_{i=1}^{n} c'_{ti} c_{it'}$$

Avec :

$$c'_{ti} = c_{it} = \begin{cases} 1 & \text{si } c(i) = t \text{ c'est-à-dire si } i \in \Pi_t \\ 0 & \text{sinon.} \end{cases}$$

On a donc que :

$$c'_{ti} c_{it'} = 1 \text{ si } i \in \Pi_t \text{ et } i \in \Pi_{t'} \text{ c'est-à-dire si } t = t'.$$

$C'HC$ sera donc bien une matrice diagonale puisque si $t \neq t'$, $(C'C)_{tt'} = 0$ et les éléments diagonaux seront donnés par :

$$(C'HC)_{tt} = \frac{1}{n}\sum_{i=1}^{n} c'_{ti} c_{it} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{c_{it} c_{it}}_{\neq 0 \text{ si } i\in\Pi_t} = \frac{1}{n}\sum_{i\in\Pi_t} \underbrace{c_{it}^2}_{=1} = \frac{1}{n}\sum_{i\in\Pi_t} c_{it} \quad \square$$

3.3.2 Matrice des centroïdes

Le centroïde de la classe C_t est donné par $\bar{x}_{C_t} = \left((\bar{x}_{C_t})_1, \ldots, (\bar{x}_{C_t})_p\right)$ où :

$$(\bar{x}_{C_t})_j = \frac{1}{n_t} \sum_{k\in C_t} x_{kj} \; ; \quad t = 1, \ldots m \; ; \; j = 1, \ldots, p.$$

Nous pouvons alors définir une matrice $\widetilde{G}$ de dimension $m \times p$ dont la t^e ligne contiendra les coordonnées du centroïde de la classe C_t :

$$\boxed{\widetilde{G} = \begin{pmatrix} \overline{x}'_{c_1} \\ \vdots \\ \overline{x}'_{c_m} \end{pmatrix} = Q^{-1}(C'HX).}$$

Preuve :

◇ $Q^{-1} = \frac{1}{detQ}\ Q^c$

$$\text{Or } (Q^c)_{tt'} = C_{t't} = (-1)^{t+t'}\ M_{t't}$$

$$\text{avec } M_{t't} = \begin{cases} 0 & \text{si } t \neq t' \text{ car Q est une matrice diagonale} \\ \frac{n_1 \ldots . n_{t-1} n_{t+1} \ldots . n_m}{n^{m-1}} & \text{si } t = t'. \end{cases}$$

$$\text{et } detQ = \frac{n_1 . \ldots . n_m}{n^m}.$$

Nous avons donc que Q^{-1} est une matrice diagonale dont les éléments diagonaux sont donnés par :

$$(Q^{-1})_{tt} = \frac{n^m}{n_1 . \ldots . n_m} \frac{n_1 \ldots . n_{t-1} n_{t+1} \ldots . n_m}{n^{m-1}} = \frac{n}{n_t}$$

◇ $(C'HX) = C'\left(\frac{1}{n} I_n\right) X = \frac{1}{n}\ C'X$

$$(C'X)_{tj} = \sum_{i=1}^{n} c'_{ti} x_{ij} = \sum_{i=1}^{n} c_{it} x_{ij} = \sum_{k \in \Pi_t} x_{kj}.$$

Puisque $c_{ki} = 0$ si $i \notin \Pi_t$. On a donc :

$$(C'HC)_{tj} = \frac{1}{n} \sum_{k \in \Pi_t} x_{kj}.$$

Pour simplifier l'écriture, nous noterons $(C'HC)_{tj}$ par y_{tj}.

◇ On obtient alors :

$$(\widetilde{G})_{ij} = \left(Q^{-1}(C'HX)\right)_{tj} = \sum_{i=1}^{n} \underbrace{Q_{ti}^{-1}}_{=0 \text{ si } i \neq t} y_{ij} = Q_{tt}^{-1} y_{ij}$$

$$= \frac{n}{n_t} . \frac{1}{n} \sum_{k \in \Pi_t} x_{kj} = \frac{1}{n_k} \sum_{k \in \Pi_t} x_{kj} = (\overline{x}_{C_t})_j . \qquad \square$$

Avec toutes ces notations, nous pouvons enfin définir les matrices de covariance inter-classes et intra-classes.

3.3.3 Matrice de covariance inter-classes

La matrice de covariance inter-classes, notée B, sera de dimension $p \times p$ et sera définie par :

$$B = \sum_{t=1}^{m} \frac{n_t}{n} \, \overline{x}_{C_t} \overline{x}'_{C_t} = (X'HC) \, Q^{-1} \, (C'HX)$$

Preuve :

$$B = (X'HC) \, Q^{-1} \, (C'HX) = (X'HC) \, \widetilde{G}$$

Or nous pouvons voir que $(X'HC) = \frac{1}{n} \, X'C$ est de dimension $p \times m$ et est de la forme :

$$\frac{1}{n} \begin{pmatrix} n_1(\overline{x}_{C_1})_1 & \dots & n_m(\overline{x}_{C_m})_1 \\ \vdots & \vdots & \vdots \\ n_1(\overline{x}_{C_1})_p & \dots & n_m(\overline{x}_{C_m})_p \end{pmatrix}.$$

Ou encore :

$$(X'HC) = \left(\frac{n_1}{n} \, \overline{x}_{C_1}, \dots, \frac{n_m}{n} \, \overline{x}_{C_m} \right).$$

Nous aurons alors que :

$$B = \Big(\frac{n_1}{n}\,\overline{x}_{C_1}, \ldots, \frac{n_m}{n}\,\overline{x}_{C_m}\Big) \begin{pmatrix} \overline{x}'_{C_1} \\ \vdots \\ \overline{x}'_{C_m} \end{pmatrix}.$$

Et donc :

$$B = \sum_{t=1}^{m} \frac{n_t}{n}\,\overline{x}_{C_t}\overline{x}'_{C_t} \qquad \square$$

Remarque :

La trace de la matrice B est la variance inter-classes des vecteurs de données $x_1, \ldots, x_n$.

Nous avons en effet qu'un élément (j,j) de la diagonale principale de la matrice B est donné par :

$$\Big(\sum_{t=1}^{m} \frac{n_t}{n}\,\overline{x}^2_{C_t}\Big)_j .$$

On a donc bien que :

$$Tr(B) = \sum_{j=1}^{p}\Big(\sum_{t=1}^{m} \frac{n_t}{n}\,\overline{x}^2_{C_t}\Big)_j$$

$$= \sum_{t=1}^{m}\Big(\sum_{j=1}^{p} \frac{n_t}{n}\,\overline{x}^2_{C_t}\Big)_j$$

$$= \text{variance inter-classes des vecteurs de données } x_1, \ldots, x_n$$
$$= I_B$$

3.3.4 Matrice de covariance intra-classes

Nous supposerons que les matrices de variance-covariance à l'intérieur des différentes classes $\Pi_1, \ldots, \Pi_m$ sont identiques.

La matrice de covariance intra-classes W est alors définie par :

$$\boxed{W = \frac{1}{n}\sum_{t=1}^{m} n_t \, W_t}$$

où W_t est la matrice de variance-covariance de la classe C_t qui est donnée par :

$$W_t = \frac{1}{n_t} \sum_{k \in C_t} (x_k - \overline{x}_{C_t})(x_k - \overline{x}_{C_t})'.$$

Remarque :

La tracc de W est la variance intra-classe des vecteurs données de E.

En effet, un élément (j, j) de la diagonale principale de la matrice W est donné par :

$$\frac{1}{n}\sum_{t=1}^{m} \sum_{k \in C_t} (x_k - \overline{x}_{C_t})_j^2 \quad j = 1, \ldots, p;$$

On a donc :

$$Tr(W) = \frac{1}{n} \sum_{j=1}^{p} \Big(\sum_{t=1}^{m} \sum_{x_k \in C_t} (x_k - \overline{x}_{C_t})_j^2 \Big)$$

$$= \frac{1}{n} \sum_{x_k \in C_t} \sum_{j=1}^{p} \Big(\sum_{t=1}^{m} (x_k - \overline{x}_{C_t})_j^2 \Big)$$

$$= \text{variance intra-classes des vecteurs de données } x_1, \ldots, x_n$$

$$= I_W$$

3.4 La méthode

3.4.1 Idée générale

Nous allons construire nos axes factoriels à partir des matrices de covariance totale et inter-classes T et B. De nouveau, nous projetterons nos données sur le sous-espace des deux premiers axes factoriels, appelé plan factoriel.

Une fois notre plan factoriel construit, nous chercherons à reclasser nos individus. Les individus seront affectés à la classe dont ils sont la plus proche. Nous devrons donc construire une règle d'affectation qui nous permette de trouver cette classe.

3.4.2 Première étape : Construction des axes factoriels

Nous cherchons s axes factoriels $F_i \in R^p \quad (i = 1, \ldots, s < p)$ tels que les groupes $C_1, \ldots, C_m$ soient aussi distincts que possible. Le critère que nous utiliserons ici est la maximisation du rapport entre l'inertie inter-classes I_B et l'inertie intra-classes I_W.

Pour trouver les axes, nous allons considérer la métrique de Mahalanobis T^{-1} sur l'espace R^p.

Distance de Mahalanobis :

Avec la distance euclidienne, nous avons que tous les points dont la distance à l'origine est identique, par exemple, $x = (x_1, \ldots, x_p)$ tel que $\| x \| = c$, satisfont $x_1^2 + \ldots x_p^2 = c^2$ qui est l'équation d'une hypersphère. Ceci signifie que toutes les composantes de l'observation x contribuent de façon égale à la distance euclidienne entre x et le centre.

Mais en statistique, chaque composante correspond à une variable et nous aimerions donc avoir une distance qui prenne en compte la variance des variables pour déterminer la distance au centre. Les composantes avec une grande variabilité devraient recevoir moins de poids que celles avec une petite variabilité. Pour cela, on pondère les composantes par un facteur a_j.

On considère alors

$$u = (\frac{x_1}{a_1}, \ldots, \frac{x_p}{a_p})$$
$$v = (\frac{y_1}{a_1}, \ldots, \frac{y_p}{a_p})$$

La distance euclidienne entre u et v devient alors :

$$d(u,v) = \sqrt{(x-y)'D^{-1}(x-y)}$$

où $D = diag(a_1^2, \ldots a_p^2)$.

Nous avons aussi que :

$$\| u \| = \sqrt{x'D^{-1}x}$$

et tous les points qui ont la même distance à l'origine satisfont :

$$(\frac{x_1}{a_1})^2 + \ldots + (\frac{x_p}{a_p})^2 = c^2$$

qui est l'équation d'un ellipsoïde centré à l'origine dont les axes sont u et v.

Nous souhaitons aussi tenir compte de la corrélation entre les variables. Pour cela, il suffit de prendre comme matrice D de poids, la matrice de variance-covariance S et nous avons alors la définition suivante pour la métrique de Mahalanobis :

$$d(x,y) = \sqrt{(x-y)'S^{-1}(x-y)}.$$

Construction des axes :

Avec cette distance, on peut alors voir que l'inertie totale dans une direction $u \in R^p$ unitaire, où F est un vecteur unitaire au sens de la métrique de Mahalanobis dans la direction u, est donnée par :

$$I_u = F'T'F = 1.$$

Or, nous avons que : $T = W + B$, ce qui permet de décomposer I_u en :

$$I_u = u'F'WF + F'BF.$$

Nous aurons donc que le pouvoir discriminant de la direction u sera d'autant plus grand que l'inertie inter-classes $F'BF$ sera grande ou, ce qui est équivalent, que l'inertie intra-classes $F'WF$ est faible.

Nous devons alors résoudre le problème d'optimisation suivant :

$$\boxed{\max F'BF \text{ sous la contrainte } F'TF = 1.}$$

Les solutions de ce problème sont obtenues par les solutions de l'équation :

$$BF = \lambda TF \text{ ou } T^{-1}BF = \lambda F.$$

Si on considère l'équation $T^{-1}BF = \lambda F$, nous pouvons voir que les solutions sont les vecteurs propres v_i de la matrice $T^{-1}B$.

De manière équivalente à l'analyse en composantes principales, nous prendrons comme axes factoriels $F_1, \ldots, F_s$ les vecteurs propres de $T^{-1}B$ qui correspondent aux s plus grandes valeurs propres $\lambda_1, \ldots, \lambda_s$ classées par ordre décroissant.

Remarque :

En général, pour que m classes soient bien séparées sur le plan factoriel et éviter que des individus soient affectés à une mauvaise classe, $m-1$ axes factoriels seront nécessaires. Ce nombre coïncide avec le nombre de valeurs propres non nulles maximum de la matrice $T^{-1}B$ et avec le rang de la matrice B.

3.4.3 Deuxième étape : Représentation des données dans l'espace de dimension s

Les coordonnées de l'individu $i \in E$ sur le plan factoriel formé par $F_1, \ldots, F_s$ sont données par :

$$y_i = \begin{pmatrix} y'_{1i} \\ \vdots \\ y'_{li} \end{pmatrix} = V'_s x_i.$$

Les coordonnées du i$^{\text{e}}$ centre sont :

$$\overline{y}_{C_i} = V'_s \, \overline{x}_{C_i}.$$

Le l$^{\text{e}}$ axe factoriel est donné par :

$$F_l = v'_l \, Y.$$

3.5 Définition d'une règle d'affectation

Nous avons maintenant obtenu un sous-espace factoriel qui optimise la séparation des classes. Nous allons reconstituer une partition des individus dans l'espace à s dimensions.

Pour cela, nous allons définir une règle d'affectation qui réaffectera les observations à une des classes $\Pi_1, \ldots, \Pi_m$. Nous comparerons ensuite la partition obtenue à la partition initiale à l'aide de cette règle.

Plusieurs règles d'affectation géométriques ont été proposées. L'une d'elle propose d'affecter un individu k à la classe Π_t si la distance de la projection de x_k sur le plan factoriel, y_k, à la projection du centroïde de la classe Π_t, y_{C_t}, est inférieure aux distances de y_k aux projections des centroïdes des autres classes.

Formellement, on a donc :

L'individu k est affecté à la classe Π_t

$$\Leftrightarrow$$

$$d(y_k, y_{C_t}) = \min_{j=1,\ldots,m} d(y_k, y_{C_j}).$$

Il nous reste alors à choisir une distance d : la distance euclidienne ou la distance de Mahalanobis.

3.6 Validation de la règle d'affectation

Comme pour l'analyse en composantes principales, nous allons essayer de trouver une mesure de la qualité des résultats de l'analyse factorielle discriminante. Les taux de bons et de mauvais classements nous permettront d'évaluer cette qualité.

3.6.1 Définitions

Le taux de bons (mauvais) classements est le pourcentage d'individus affectés à une bonne classe (à une mauvaise classe).

Ces taux doivent être calculés sur l'ensemble E et sur chacune des classes.

On a par exemple dans le cas de 2 classes :

		Affectation		
	Classes	C_1	C_2	Nombres total d'individus
Réalité	C_1	a	b	N_1
	C_2	c	d	N_2

où a, b, c, d, N_1 et N_2 sont des entiers positifs.

Les taux de bons classements sont donnés par :

$\rightarrow$ Pour la classe C_1 : $t_1 = \frac{a}{N_1} \quad 0 \leq t_1 \leq 1$

$\rightarrow$ Pour la classe C_2 : $t_2 = \frac{d}{N_2} \quad 0 \leq t_2 \leq 1$

$\rightarrow$ Pour l'ensemble E : $t = \frac{a+d}{N_1+N_2} = \frac{N_1 t_1 + N_2 t_2}{N_1+N_2} \quad 0 \leq t \leq 1$

Et les taux de mauvais classements par :

$\rightarrow$ Pour la classe C_1 : $e_1 = \frac{b}{N_1} \quad 0 \leq t_1 \leq 1$

$\rightarrow$ Pour la classe C_2 : $e_2 = \frac{c}{N_2} \quad 0 \leq t_2 \leq 1$

$\rightarrow$ Pour l'ensemble E : $e = \frac{b+c}{N_1+N_2} = \frac{N_1 e_1 + N_2 e_2}{N_1+N_2} \quad 0 \leq t \leq 1.$

On peut alors voir pourquoi il faut tenir compte des taux des classes et de ceux de l'ensemble E. En effet, si on a, par exemple, un mauvais taux de bons classements pour un groupe (t_i petit) qui ne contient pas beaucoup d'éléments (N_i petit), le taux global ne sera pas fort influencé par ce mauvais taux et on pourrait alors tirer de mauvaises conclusions sur le taux de bons classements.

Les taux de bons et de mauvais classements calculés sur l'ensemble E sont appelés les taux calculés par *resubstitution.*

Mais nous souhaitons savoir si la règle d'affectation que nous avons définie nous permettait de classer correctement de *nouveaux individus.* Il nous faut donc définir les taux *réels* (ou théoriques) de bons et de mauvais classements.

3.6.2 Taux réels de bons et de mauvais classements

Pour les taux sur l'ensemble test :

Le *taux réel de bons classements* $TBC^*(d)$ est la probabilité que la règle de décision d classe correctement les individus d'un nouvel échantillon tiré de la même population totale.
Le *taux réel de mauvais classements* $R^*(d)$ est la probabilité que la règle de décision d classe mal les individus d'un nouvel échantillon tiré de la même population totale.

Les taux calculés par resubstitution sous-estiment $R^*(d)$ et surestiment $TBC^*(d)$ puisque la règle de décision d peut être influencée par certaines particularités de l'ensemble E surtout lorsque E est petit.

Pour les taux sur chacun des groupes :

Supposons dans le cas de 2 groupes C_1 et C_2 que les taux de bons classements réels soient $\widehat{t_1}$ et $\widehat{t_2}$. On a alors que le taux de bons classements global est :

$$\widehat{t} = \pi_1 \widehat{t_1} + \pi_2 \widehat{t_2}$$

où π_i est la probabilité d'appartenance au groupe i a priori.

Les taux de mauvais classements réels $\widehat{e_1}$ et $\widehat{e_2}$ sont :

$$\widehat{e_1} = 1 - \widehat{t_1} \text{ et } \widehat{e_2} = 1 - \widehat{t_2}.$$

Ce sont ces taux $\widehat{t_i}$ et $\widehat{e_i}$ que nous allons chercher à estimer pour avoir une mesure de la qualité des résultats que l'on obtient avec l'analyse factorielle discriminante.

Pour les évaluer, nous verrons 3 méthodes :

- les échantillons tests
- le leave-one-out et
- le bootstrap.

3.6.3 Evaluation des taux réels :

Echantillon test :

Puisque nous essayons de voir si notre règle d'affectation d nous permet de classer de nouveaux individus, nous pouvons considérer un nouvel ensemble d'individus, appelé ensemble test, sur lequel nous appliquerons d. Nous pourrons alors calculer le taux de mauvais classements de cet ensemble qui sera généralement plus élevé que celui calculé par resubstitution : ce taux approche en effet le taux réel.

Cet échantillon peut être obtenu de deux manières différentes :

- Si nous disposons d'un ensemble de N individus différents sur lesquels les mêmes variables Y_j ont été mesurées, il peut alors être considéré comme ensemble test que nous noterons $E^+ = \{1, \ldots, N\}$.

- Si nous ne disposons pas de données différentes, on prend un sous-ensemble de E, de dimension N $(N < n)$ comme ensemble test E^+. Mais de cette façon, on se prive d'une partie des informations contenues dans les données pour estimer les taux.

Méthode du Leave-one-out

Nous allons considérer successivement les N ensembles

$$E_v = E - L_v \ \ (v = 1, \ldots, N).$$

Sur chacun de ces ensembles, on appliquera la règle de décision d. On notera par $d^{(v)}$ la règle de décision lorsque celle-ci est appliquée à l'ensemble E_v.

A chaque fois, nous calculerons $R^{*v}(d^{(v)})$ le taux de mauvais classements de l'ensemble E_v. La moyenne de ces N taux de mauvais classements nous donne le taux de mauvais classements réel :

$$R^* = \frac{1}{N} \sum_{v=1}^{N} R^{*v}(d^{(v)}).$$

Méthode du Bootstrap

Pour cette méthode, nous allons effectuer B tirages de n éléments avec remise parmi les n individus de l'ensemble E pour construire B nouveaux échantillons appelés *échantillons bootstrap*.

Chaque individu a une probabilité $\frac{1}{n}$ d'être tiré. Puisque le tirage s'effectue *avec remise*, certains individus seront tirés plusieurs fois et d'autres pas du tout.

On calcule, pour chacun des B échantillons bootstrap, les pourcentages de mauvais classements que nous noterons, pour le b^{e} échantillon, $\widehat{R}^{*b}$ ($b = 1, \ldots, B$). Nous pourrons alors estimer le taux de mauvais classements par la moyenne des taux obtenus sur les échantillons bootstrap :

$$R^* = \frac{1}{B} \sum_{b=1}^{B} \widehat{R}^{*b}.$$

3.7 Exemples

3.7.1 Sur des données artificielles

Soit 10 objets sur lesquels on mesure 4 variables quantitatives réelles :

	Y_1	Y_2	Y_3	Y_4
1	2	-7	-1	-1
2	-5	3	-15	5
3	4	-4	11	-2
4	0	2	6	1
5	0	1	0	-3
6	9	5	7	0
7	-10	-2	1	0
8	-5	0	-4	1
9	-1	-5	-5	4
10	6	7	0	-5

Commençons par rappeler les différentes étapes de l'analyse factorielle discriminante :

1. On calcule les différentes matrices dont nous allons avoir besoin :

$$X, H, C, Q, T \text{ et } B$$

2. On calcule $T^{-1}B$ ainsi que ses valeurs propres et ses vecteurs propres.
3. On obtient les nouvelles coordonnées des individus sur le plan factoriel par

$$y_i = V_s' x_i$$

où V_s' est la matrice contenant, en colonnes, les vecteurs propres que nous avons retenus pour construire les axes factoriels.
4. On affecte les individus à la classe dont ils sont la plus proche c'est-à-dire à la classe C_t telle que :

$$d(y_k, y_{C_t}) = \min_{j=1,\ldots,m} d(y_k, y_{C_j})$$

5. On valide les résultats obtenus soit par un échantillon test, soit par la méthode du bootstrap ou encore par la méthode du leave-one-out.

Première étape : Définitions des différentes matrices

Matrice d'appartenance aux différentes classes C :

Les objets appartiennent à 3 classes C_1, C_2 et C_3 suivant la valeur de la variable Y_1.

Si :

- $Y_1 > 0$: l'objet appartient à C_1
- $Y_1 < 0$: il appartient à C_2
- $Y_1 = 0$: il appartient à C_3.

On obtient alors la matrice suivante :

$$C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Matrice de covariance T :

$$T = X'HX$$

où H est la matrice des poids des objets :

$$H = \frac{1}{10} I_{10}.$$

Nous avons alors :

$$T = \begin{pmatrix} 28.8 & 6.7 & 19.5 & -7.4 \\ 6.7 & 18.2 & -1.2 & -2.6 \\ 19.5 & -1.2 & 47.4 & -11.4 \\ -7.4 & -2.6 & -11.4 & 8.2 \end{pmatrix}.$$

Matrice des poids des classes Q :

$$Q = C'HC$$

$$= \begin{pmatrix} 0.4 & 0 & 0 \\ 0 & 0.4 & 0 \\ 0 & 0 & 0.2 \end{pmatrix}.$$

Matrice des centroïdes $\widetilde{G}$:

$$\widetilde{G} = Q^{-1}(C'HX)$$

$$= \begin{pmatrix} 5.2 & 0.25 & 4.25 & -2 \\ -5.25 & -1 & -5.75 & 2.5 \\ 0 & 1.5 & 3 & -1 \end{pmatrix}.$$

Matrice de covariance inter-classes B :

$$B = (X'HC)Q^{-1}(C'HX)$$

$$= \begin{pmatrix} 22.5 & 2.625 & 21 & -9.45 \\ 2.625 & 0.875 & 3.625 & -1.5 \\ 21 & 3.625 & 22.25 & -9.75 \\ -9.45 & -1.5 & -9.75 & 4.3 \end{pmatrix}.$$

Deuxième étape : Construction des axes factoriels

Nous devons donc calculer les valeurs propres et les vecteurs propres de $T^{-1}B$ c'est-à-dire de :

$$T^{-1}B = \begin{pmatrix} 0.6273 & 0.0345 & 0.5172 & -0.2421 \\ -0.1673 & -0.0262 & -0.0672 & 0.041 \\ 0.0405 & 0.0427 & 0.1143 & -0.0426 \\ -5831 & -0.0841 & -0.5847 & 0.2597 \end{pmatrix}.$$

On trouve alors :

$$\lambda_1 = 0.8989$$
$$\lambda_2 = 0.1286$$

$$\lambda_3 = 8.6221.10^{-17}$$
$$\lambda_4 = -3.4149.10^{-17}.$$

En calculant l'inertie totale (1.0275) et les pourcentages de variance expliqués par chaque valeur propre, nous pouvons voir que deux axes factoriels nous suffisent à expliquer la presque totalité de cette variance :

$$\frac{\lambda_1 + \lambda_2}{I} \cong 1.$$

Les vecteurs propres correspondants sont :

$$v_1 = \begin{pmatrix} -0.708 \\ 0.1727 \\ -0.0642 \\ 0.6818 \end{pmatrix} ; \quad v_2 = \begin{pmatrix} -0.6036 \\ 0.6207 \\ 0.487 \\ -0.1145 \end{pmatrix}.$$

Troisième étape : Représentation des données sur le plan factoriel

Les coordonnées des points sont données par :

$$y_i = V_s' x_i \; ; \quad i = 1, \ldots, 10$$

où $V's$ est la matrice contenant v_1 et v_2.

On obtient alors :

i	y_{i1}	y_{i2}
1	-3.2425	-5.925
2	8.4294	-2.29978
3	-5.5921	0.689
4	0.6422	4.0492
5	-1.8726	0.9643

6	-5.9574	1.0803
7	6.6702	5.2818
8	4.4783	0.9555
9	2.8924	-5.3934
10	-6.4478	1.2961

Quatrième étape : Définition d'une règle d'affectation

Pour cela, calculons les coordonnées des centres des trois classes sur le plan factoriel :

$$y_{C_1} = \begin{pmatrix} -5.31 \\ -0.7149 \end{pmatrix}; \quad y_{C_2} = \begin{pmatrix} 5.6176 \\ -0.7149 \end{pmatrix}; \quad y_{C_3} = \begin{pmatrix} -0.6152 \\ 2.5067 \end{pmatrix}.$$

Notre règle est d'affecter l'individu i à la classe C_t si :

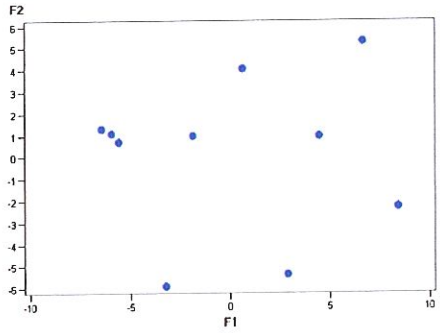
$$d(y_{C_t}, y_i) = \min_{t=1,2,3} d(y_{C_t}, y_i).$$

Calculons ensuite la distance de chaque objet aux centroïdes des différentes classes :

	y_{C_1}	y_{C_2}	y_{C_3}
y_1	5.6053	10.369	8.8315
y_2	13.9278	3.7356	10.5879
y_3	1.432	11.267	5.2984
y_4	7.624	6.7677	1.9901
y_5	3.8256	7.6395	1.99
y_6	1.9084	11.6876	5.5294
y_7	13.3972	5.9147	7.796
y_8	9.9298	1.8788	5.3245
y_9	8.9181	5.8901	8.4147
y_10	2.3106	12.2041	5.9569

Les individus $1, 3, 6$ et 10 sont affectés à la classe C_1, $2, 7, 8$ et 9 à la classe C_2 et les individus 4 et 5 à C_3. Tous les individus se retrouvent donc dans leurs classes connues a priori.

On a la représentation suivante :



Cinquième étape : Validation

La méthode utilisée ici est celle de l'échantillon test. Soit 6 individus t_N sur lesquels les mêmes variables ont été observées :

	Y_1	Y_2	Y_3	Y_4
t_1	-1	2	1	3
t_2	-6	2	1	2
t_3	0	3	0	-2
t_4	5	6	4	-1
t_5	2	2	-4	-3
t_6	0	4	-2	1

A priori, les individus t_4, t_5 appartiennent à C_1, t_1, t_2 à C_2 et t_3 et t_6 à C_3.

On calcule alors les coordonnées de ces points sur le plan factoriel :

$$z_{t1} = \begin{pmatrix} 3.0346 \\ 1.9886 \end{pmatrix}; \quad z_{t2} = \begin{pmatrix} 5.8927 \\ 5.1212 \end{pmatrix}; \quad z_{t3} = \begin{pmatrix} -0.8455 \\ 2.0913 \end{pmatrix};$$

$$z_{t4} = \begin{pmatrix} -3.4421 \\ 2.769 \end{pmatrix}; \quad z_{t5} = \begin{pmatrix} -2.8592 \\ -1.5703 \end{pmatrix}; \quad z_{t6} = \begin{pmatrix} 1.5009 \\ 1.3944 \end{pmatrix}.$$

On teste alors notre règle d'affectation sur ces nouveaux individus :

	y_{C_1}	y_{C_2}	y_{C_3}
t_1	8.7715	3.6136	3.6864
t_2	12.6317	5.6664	7.0134
t_3	5.2732	6.9776	7.0134
t_4	3.953	9.6446	2.839
t_5	2.5958	8.5394	4.6538
t_6	7.13	4.5478	2.3907

t_1 est donc affecté à C_1, t_1 et t_2 à C_2 et t_3, t_4 et t_6 à C_3.

Nous pouvons alors voir que le taux de mauvais classement est de $1/6$.

3.7.2 Sur des données réelles : Les Iris de Fisher

FIG. 3.1 – Iris Setosa, Versicolor et Virginica (The species iris group of north America, http ://www.badbear.com/signa/signa.pl)

Présentation des données

Ces données ont été proposées en 1933 par Ronald Aylmer Fisher. Elles comportent la description de 150 iris de trois espèces différentes :

- 50 setosa
- 50 versicolor
- 50 virginica

par quatre variables :

- Y_1 : la longueur du sépale
- Y_2 : la largeur du sépale
- Y_3 : la longueur des pétales

- Y_4 : la largeur des pétales.

Ces 150 iris sont logiquement répartis en 3 classes qui correspondent aux 3 espèces d'iris :

- C_1 : les iris setosa
- C_2 : les iris versicolor
- C_3 : les iris virginica.

Chacune de ces classes comporte 50 individus.

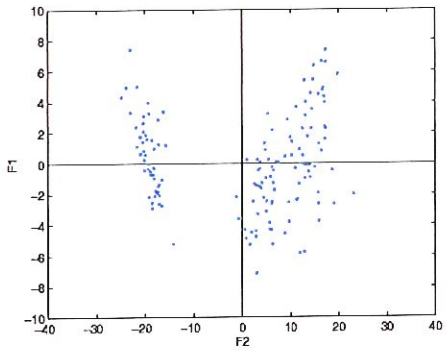
Ces données sont disponibles sur *http ://www.info.univ-anger.fr/ gh/Datasets/iris.htm.*

Analyse factorielle discriminante classique :

Nous pouvons ramener ces variables sur un plan factoriel formé par :

$$F_1 = -0.2087Y_1 - 0.3862Y_2 + 0.554Y_3 + 0.7074Y_4$$
$$F_2 = 0.0065Y_1 + 0.5866Y_2 - 0.2526Y_3 + 0.7695Y_4$$

Nous obtenons alors la représentation suivante :



Règle d'affectation :

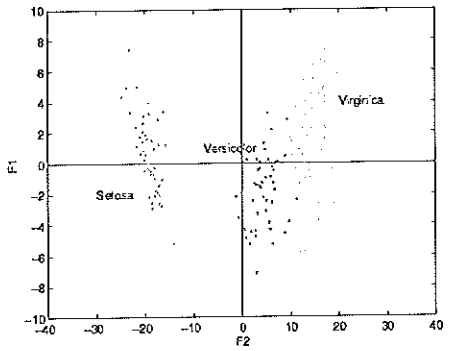
Maintenant que nous avons représenté les iris sur le plan factoriel, regardons à quelle classe ils sont affectés.

Notre règle d'affectation étant donnée par :
l'individu i est affecté à C_t si :

$$d(y_{C_t}, y_i) = \min_{t=1,2,3} d(y_{C_t}, y_i)$$

où C_t représente le centre de la classe t $(t = 1, 2, 3)$.

Nous obtenons alors les classes suivantes :



Les taux e_t de mauvais classements par classe sont :

$$e_1 = 0$$

$$e_2 = \frac{2}{150} = 0.013$$

$$e_3 = \frac{1}{150} = 0.006.$$

Et le taux de mauvais classements général e :

$$e = \frac{3}{150} = 0.02$$

Validation :

Nous allons ici utiliser la méthode du bootstrap. Pour cela, nous prenons au hasard 150 individus parmi les individus de départ. Si nous formons 3 échantillons bootstrap B_v nous obtenons :

- Pour B_1 : $\widehat{R}^{*1} = 0$
- Pour B_2 : $\widehat{R}^{*2} = 0.026$
- Pour B_3 : $\widehat{R}^{*3} = 0.02$.

La moyenne de ces taux de mauvais classements nous donne le taux de mauvais classements réel :

$$R^* = \frac{1}{3}\sum_{v=1}^{3} \widehat{R}^{*v} = 0.015.$$

Comme nous pouvons le voir sur le graphique, le problème pour classer les iris sur le plan factoriel vient des classes iris versicolor et virginica puisque ces classes sont très proches.

Chapitre 4

Analyse factorielle discriminante symbolique

4.1 Introduction

Nous allons généraliser l'analyse factorielle discriminante classique aux données symboliques.

De manière similaire à ce que nous avons fait pour l'analyse en composantes principales pour des données intervalles, nous allons nous ramener à une application de l'analyse factorielle discriminante classique.

Pour pouvoir travailler avec tous les types de données symboliques simultanément, nous allons commencer par leur donner une certaine homogénéité en codant les variables. Nous pourrons , à partir de ces variables codées, obtenir une représentation géométrique des objets symboliques par des hyperrectangles.

Le codage nous fera perdre la compacité des objets, nous la retrouverons en quantifiant les variables codées représentant les sommets appartenant aux différents objets. Nous appliquerons l'analyse factorielle discriminante classique sur ces variables quantifiées. Nous pourrons alors reconstituer les données symboliques sur le plan factoriel par le rectangle d'aire maximum couverte par les sommets des hyperrectangles.

Nous construirons ensuite une règle d'affectation et les méthodes classiques de validation resteront applicables ici.

Nous terminerons ce chapitre par un exemple de données artificielles et par un exemple de données réelles.

4.2 Représentation des données

De façon similaire à la méthode classique, on part avec :

- n objets u appartenant à un ensemble $E = \{1, \ldots, n\}$ également appelé ensemble d'apprentissage ou de base
- $Y_1, \ldots, Y_p$, p variables *symboliques* de domaine $\mathcal{Y}_1, \ldots, \mathcal{Y}_p$ respectivement
- n données $x_1, \ldots, x_n$ où $x_u = (x_{u1}, \ldots, x_{up})'$ avec $x_{uj} = Y_j(u)$
- m classes $\Pi_1, \ldots, \Pi_m$ auxquelles appartiennent les différents objets.

Suivant le type de variable symbolique auquel nous avons affaire, le domaine $\mathcal{Y}_j$ sera :

- $\mathbb{R}$ pour une variable quantitative
- l'ensemble des intervalles fermés bornés de $\mathbb{R}$ pour une variable intervalle
- un ensemble de catégories pour une variable qualitative (nominale ou ordinale) ou une variable multivaluée
- un ensemble de catégories auxquelles nous associerons des fréquences pour une variable modale.

Ici aussi, on note par C_t l'ensemble des éléments de E qui appartiennent à la classe Π_t :

$$C_t = \{u \in E | u \in \Pi_t\}.$$

Nous pouvons décrire C_t (et donc Π_t) par un objet symbolique de type :

$$D_t = D_{t1} \times \ldots \times D_{tj} \ldots \times D_{tp}$$

où chaque D_{tj} est un sous-ensemble du domaine $\mathcal{Y}_j$. On a donc :

$$D_t \in \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_p.$$

Plus précisément, nous supposerons que chaque domaine $\mathcal{Y}_j$ a été partitionné en un nombre d'intervalles ou de sous-ensembles et D_{tj} est un sous-ensemble de la partition résultante $P(\mathcal{Y}_j)$: $D_{tj} \in P(\mathcal{Y}_j)$.

Nous devons ensuite pouvoir comparer la valeur réalisée de l'individu u, $Y_j(u)$, à D_{tj}. Pour cela, on considère p relations $\mathcal{R}_j$ $(j = 1, \ldots, p)$ et on définit la fonction boolénne q_t (pour $t = 1, \ldots, m$) par :

$$q_i(u) = \begin{cases} 1 & \text{si } Y_j(u)\ \mathcal{R}_j\ D_{ij} \text{ pour } j = 1, \ldots, p \\ 0 & \text{sinon.} \end{cases}$$

Les éléments $u \in E$ tels que $q_t(u) = 1$ satisfont donc les p relations. Ces éléments font partie de l'extension des objets symboliques dans E qui décrit la classe C_t (ou Π_t) :

$$Ext(C_t|E) = \{u \in E | q_t(u) = 1\}.$$

De cette manière, une classe C_t correspond à un objet symbolique du deuxième ordre c'est-à-dire à un ensemble de classes d'individus.

Cependant, nous considèrerons ici seulement des objets symboliques du premier ordre et l'appartenance de chaque objet de l'ensemble E sera définie par une variable qualitative c :

$$c(u) = t \text{ si } u \in \Pi_t.$$

Ou encore par p variables binaires :

$$c_t(u) = \begin{cases} 1 & \text{si } u \in \Pi_t \\ 0 & \text{sinon.} \end{cases}$$

4.3 La méthode

4.3.1 Codage des variables symboliques

Les variables symboliques avec lesquelles nous allons travailler pourront être de différents types : quantitatives, qualitatives, multivaluées, modales et intervalles.
Pour pouvoir travailler avec ces variables simultanément, nous allons chercher à leur donner une certaine homogénéité en **codant** ces variables suivant leur type.

Les valeurs de codage de la variable Y_j obtenues pour chaque objet seront placées dans une matrice de codage notée X_j.

Pour des variables qualitatives ou multivaluées :

Pour ce type de variable symbolique, nous appliquerons un codage dit binaire ou disjonctif complet c'est-à-dire que les valeurs de codage seront 0 ou 1 suivant que les différentes catégories se situent ou non dans la description de l'objet.

La matrice de codage X_j aura autant de colonnes que le nombre de catégories de la variable explicative Y_j et autant de lignes que de catégories intervenant dans la description des différents objets.

Exemple :

Soit la variable Y_1 représentant le genre d'un film. Cette variable est décrite par 7 modalités : catastrophe, drame, thriller, action, science-fiction, comédie et famille.
Soit 5 films u, v, w, x et z. Le film u est une catastrophe dramatique, v un thriller, w est un comédie et un film d'action et de science-fiction, x une comédie et z est un film familial.
Nous aurons alors la matrice de codage X_1 suivante :

$$X_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Pour des variables modales :

Dans ce cas, on utilise les modes associés à la variable modale Y_j comme valeurs de codage. X_j sera alors une matrice ligne qui contiendra les valeurs des modes.

Exemple :

Supposons qu'à la variable Y_1 nous associons la variable Y_2 décrivant la probabilité de choisir un film appartenant aux différentes catégories de Y_1. Si nous supposons que ces choix sont équiprobables, nous avons alors :

$$X_2 = (1/7\ 1/7\ 1/7\ 1/7\ 1/7\ 1/7\ 1/7).$$

Pour des variables quantitatives :

Nous utiliserons ici un système de codage flou pour préserver autant que possible l'information numérique contenue dans les variables originales.
Dans ce but, une variable quantitative sera transformée en une variable qualitative en utilisant une fonction polynomiale comme les B-splines.
Afin d'obtenir un nombre raisonnable de catégories pour les variables codées, on utilisera généralement des polynômes de degrés faibles comme une fonction B-spline de degré 1.

Pour cela, divisons le domaine $\mathcal{Y}_j$ de chaque variable Y_j en deux intervalles quelconques. Nous obtenons 3 noeuds que nous noterons x_i : la borne inférieure du premier sous-intervalle (x_0), la valeur de coupure du domaine (x_1) et la borne supérieure du second sous-intervalle (x_2).

Les trois fonctions B-splines sont alors définies de la façon suivante :

$$B_a(x) = \begin{cases} 0 & \text{si } x \leq x_{i-1} \\ \frac{x-x_{i-1}}{x_i-x_{i-1}} & \text{si } x_{i-1} \leq x \leq x_i \\ \frac{x_{i+1}-x}{x_{i+1}-x_i} & \text{si } x_i \leq x \leq x_{i+1} \\ 0 & \text{si } x \geq x_{i+1}. \end{cases}$$

où $a = 1, 2, 3$.

La matrice de codage X_j aura donc 3 colonnes et 1 ligne par objet que l'on souhaite coder :

$$X_j = (B_1(Y_j(u)), B_2(Y_j(u))B_3(Y_j(u))).$$

Exemple :

Soit Y_3 une variable donnant le bénéfice moyen effectué par la location d'un film. $\mathcal{Y}_3$ est donc $\mathbb{R}^+$.

Nous avons les valeurs suivantes :

$$\begin{aligned} Y_3(u) &= 500 \\ Y_3(v) &= 120 \\ Y_3(w) &= 300 \\ Y_3(x) &= 200 \\ Y_3(z) &= 600. \end{aligned}$$

On obtient alors la matrice de codage suivante :

$$X_3 = \begin{pmatrix} B_1(Y_j(u)) & B_2(Y_j(u)) & B_3(Y_j(u)) \\ B_1(Y_j(v)) & B_2(Y_j(v)) & B_3(Y_j(v)) \\ B_1(Y_j(w)) & B_2(Y_j(w)) & B_3(Y_j(w)) \\ B_1(Y_j(x)) & B_2(Y_j(x)) & B_3(Y_j(x)) \\ B_1(Y_j(z)) & B_2(Y_j(z)) & B_3(Y_j(z)) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0.6 & 0.4 & 0 \\ 0 & 1 & 0 \\ 1/3 & 2/3 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Pour des variables intervalles :

Ici aussi les variables seront codées de manière floue au moyen de trois fonctions B-splines.

On applique alors les 3 fonctions aux bornes de l'intervalle représentant $Y_j(u)$.

La matrice de codage X_j aura donc 3 colonnes et 2 lignes par objet symbolique u sur lequel on mesure Y_j :

$$X_j = \begin{pmatrix} B_1(\underline{x_{uj}}) & B_2(\underline{x_{uj}}) & B_3(\underline{x_{uj}}) \\ \\ B_1(\overline{x_{uj}}) & B_2(\overline{x_{uj}}) & B_3(\overline{x_{uj}}) \end{pmatrix}.$$

Exemple :

Soit Y_4 une variable de type intervalle décrivant le nombre moyen de locations des films u, v, w, x et z par semaine. On sait que le nombre maximum de locations par semaine pour chacun de ces films est de 16. Le domaine de Y_4, $\mathcal{Y}_4$, est donc $[0, 16]$.

Nous avons les données suivantes :

$$\begin{aligned} Y_4(u) &= [2,3] \\ Y_4(v) &= [3,9] \\ Y_4(w) &= [2,10] \\ Y_4(x) &= [9,10] \\ Y_4(z) &= [2,4] \end{aligned}$$

On divise alors $\mathcal{Y}_4$ en 2 sous-intervalles et nous obtenons $[0,8]$ et $[8,16]$ c'est-à-dire : $x_0 = 0, x_1 = 8$ et $x_2 = 16$. Il nous reste alors à calculer les valeurs des 3 fonctions B-splines aux bornes supérieures et inférieures des données.

On obtient alors la matrice de codage suivante :

$$X_4 = \begin{pmatrix} 3/4 & 1/4 & 0 \\ 5/8 & 3/8 & 0 \\ 5/8 & 3/8 & 0 \\ 0 & 7/8 & 1/8 \\ 2/4 & 1/4 & 0 \\ 0 & 3/4 & 1/4 \\ 0 & 7/8 & 1/8 \\ 0 & 3/4 & 1/4 \\ 3/4 & 1/4 & 0 \\ 1/2 & 1/2 & 0 \end{pmatrix}.$$

Remarque :

- Nous pouvons voir que dans le cas du codage flou, la somme des valeurs de chaque ligne est égale à 1.

- Les valeurs des différentes fonctions B-splines peuvent être interpretées comme le degré d'appartenance des différents objets aux différents niveaux de la variable Y_j : bas, moyen et haut. Dans notre exemple, nous aurons que la première fonction représentera un faible nombre de locations, la deuxième un nombre moyen de locations et la troisième un nombre élevé de locations.

4.3.2 Représentation matricielle des objets symboliques

Nous pouvons alors construire une matrice de codage globale X. Cette matrice de codage peut être construite en combinant toutes les catégories présentes dans les p variables symboliques ou encore en juxtaposant les p matrices de codage X_j :

$$X = [X_1|\ldots|X_j|\ldots|X_p].$$

▷ **Le nombre K de colonnes** de la matrice est le nombre de modalités présentes dans les p variables symboliques.
Pour un codage binaire, on a autant de colonnes que de modalités de la variable et pour un codage flou, nous avons toujours 3 colonnes. Ces nombres doivent être additionnés puisque, dans la matrice X, nous combinons toutes les possibilités.

On a alors :

$$\boxed{K = 3g + \sum_{j=1}^{q} k_j}$$

où :

- g ($\leq p$) est le nombre de variables quantitatives et intervalles sur lesquelles on applique un codage flou
- $q = p - g$ est le nombre de variables nominales, qualitatives et modales
- k_j est le nombre de catégories de la variable Y_j ($j = 1, \ldots, q$).

Exemple :

Si nous reprenons nos variables Y_1, Y_2, Y_3 et Y_4 et nos films u, v, w, x et z nous obtenons :

$$g = 2$$
$$q = 4 - 2 = 2$$
$$k_j = 7\ j = 1, 2$$
$$\text{Donc, } K = 6 + 7 + 7 = 20.$$

▷ **Le nombre N de lignes** de X sera plus grand que le nombre n d'objets dans l'ensemble E puisque chaque objet est codé sur *au moins* une ligne.

Pour un codage binaire, nous avons autant de lignes que de modalités présentes dans la description de l'objet dans le cas des variables nominales et 1 ligne pour des variables modales. Pour un codage flou, dans le cas de variables quantitatives, on a toujours 1 ligne et dans le cas des variables intervalles, à chaque objet correspond 2 lignes.

Ces nombres doivent être multipliés pour chaque objet puisque l'on combine toutes les possibilités. Nous aurons en effet que, pour un objet symbolique u, à chaque modalité présente dans sa description, on associera les différentes combinaisons possibles pour les modalités des autres variables.

Nous aurons donc :

$$N = \sum_{u=1}^{n} [2^h \prod_{v=1}^{l} k_{uj_v}]$$

où :

- h ($\leq g$) est le nombre de variables intervalles
- $l = q + (g - h)$ est le nombre de variables multivaluées (catégoriques ou quantitatives). Elles sont notées par $y_{j_1}, \ldots, y_{j_l}$
- $k_{uj} = |Y_j(u)|$ est le nombre de catégories de la variable nominale Y_j ou le nombre de valeurs de la variable multivaluée quantitative Y_j pour l'objet symbolique u.

Exemple :

$$h = 2$$

$$l = 2 + (2 - 2) = 2$$

$$k_{uj} = k_{wj} = 2$$

$$k_{vj} = 1$$

Donc, $N = 2.(2.1) + 2.(1.1) + 2.(2.1) + 2.(1.1) + 2.(1.1) = 14.$

On obtient ainsi la matrice de codage X suivante :

$$\left(\begin{array}{cccccccccccccccccc}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 3/4 & 1/4 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 5/8 & 3/8 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 3/4 & 1/4 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 5/8 & 3/8 & 0 \\
 & & & & & & & & & & & & & & & & & & & \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0.6 & 0.4 & 0 & 5/8 & 3/8 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0.6 & 0.4 & 0 & 0 & 7/8 & 1/8 \\
 & & & & & & & & & & & & & & & & & & & \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 3/4 & 1/4 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 0 & 3/4 & 1/4 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 3/4 & 1/4 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 0 & 3/4 & 1/4 \\
 & & & & & & & & & & & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/3 & 2/3 & 0 & 0 & 7/8 & 1/8 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/3 & 2/3 & 0 & 0 & 3/4 & 1/4 \\
 & & & & & & & & & & & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 3/4 & 1/4 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 1/2 & 1/2 & 0
\end{array}\right)$$

Remarque :

Les variables modales n'augmentent pas le nombre de lignes de X puisqu'elles sont codées en une seule ligne pour chaque objet symbolique par les valeurs de leurs modes.

Une fois la matrice de codage générale construite, nous pouvons également construire une matrice G de dimension $N \times n$ dont les éléments seront 0 ou 1 suivant que la ligne correspondante de X appartient aux n objets symboliques :

$$G_{ij} = \begin{cases} 1 & \text{si la i}^{\text{e}} \text{ ligne de } X \text{ appartient au j}^{\text{e}} \text{ objet symbolique} \\ 0 & \text{sinon.} \end{cases}$$

Exemple :

Dans notre exemple, on a :

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

4.3.3 Représentation géométrique des objets symboliques codés

Le codage nous permet d'analyser les objets symboliques par des méthodes numériques classiques, mais il nous fait perdre la compacité des objets symboliques. En effet, au lieu d'avoir, par exemple, deux catégories pour décrire un objet, nous avons une matrice de dimensions $2 \times k_j$ où k_j est le nombre de catégories de Y_j. L'interprétation doit donc nous permettre de retrouver l'information relative à chaque objet symbolique. Généralement, ces informations seront retrouvées à partir d'une représentation graphique.

Comme nous l'avons fait pour l'analyse en composantes principales intervalle, nous pouvons représenter chaque ligne de la matrice de codage X par un sommet. Un objet symbolique u sera alors représenté par un hyperrectangle à K sommets où K est le nombre de colonnes de X par un ensemble de sommets. Un objet symbolique u sera donc représenté par :

$$\left(2^h \prod_{v=1}^{l} k_{uj_v}\right) \text{ sommets.}$$

Pour reconstituer un objet symbolique u à partir de ses sommets, on considère son rectangle d'aire maximum couverte ($mcar$). Les côtés de ce rectangle, parallèles aux axes, passent par les sommets dont les coordonnées sont maximales et minimales. Ce rectangle contient donc tous les sommets représentant l'objet symbolique u.

4.3.4 Quantification des variables symboliques sous contraintes

Nous avons obtenu un ensemble de variables homogènes grâce à un codage binaire et à un codage flou. Pour prendre en compte la cohésion parmi les sommets appartenant au même objet symbolique, nous allons à présent quantifier ces variables codées. Pour cela, nous allons considérer la généralisation de l'analyse en composantes principales classique valable pour tous les types de variables symboliques : l'analyse canonique généralisée symbolique (SGCA).

Principe de l'analyse canonique généralisée symbolique :

Tout comme l'analyse factorielle discriminante symbolique, cette méthode part d'un ensemble de variables codées et d'une matrice de codage globale X. Le critère qu'elle cherche à optimiser est l'inertie totale des sommets appartenant aux différents objets symboliques en tenant compte de la cohésion entre les sommets appartenant à un même objet symbolique.

Pour cela, nous allons projeter les hyperrectangles sur les sous-espaces E_j formés par les vecteurs colonnes de la matrice de codage de la variable Y_j, X_j. On cherchera des axes de synthèses dans chacun de ces sous-espaces et des vecteurs orthogonaux ψ_r qui seront une synthèse de tous ces axes. Le nombre de vecteurs ψ_r est donné par $K - p + 1$.

Ces vecteurs ψ_r seront obtenus par les vecteurs propres de la matrice de dimension $n \times n$ suivante :

$$\frac{1}{N} G' X \Delta_x^I X' G$$

où Δ_x^I est une matrice diagonale par blocs de dimension $K \times K$ où les différents blocs sont $(X_j' X_j)^{-1}$.

Nos variables quantifiées que nous noterons Φ_r correspondent aux r^e coordonnées des sommets sur le plan formé par ces vecteurs propres (généralement au nombre de $K - p + 1$). Elles sont donc données par les colonnes de la matrice de dimensions $N \times (K - p + 1)$ suivante :

$$\widehat{X} = X \Delta_x^I X' G V$$

où V est la matrice dont les colonnes sont les vecteurs propres ψ_r.

4.3.5 Application de l'analyse factorielle discriminante classique

Dans le même ordre d'idée que pour l'analyse en composantes principales symboliques, nous allons appliquer l'analyse factorielle discriminante classique aux variables symboliques quantifiées que nous venons de construire.

Pour cela, considérons une matrice indicatrice C qui indique l'appartenance de chacun des N sommets à une des m classes C_t.

$$C_{it} = \begin{cases} 0 & \text{si l'individu auquel appartient le sommet } i \text{ appartient à la classe } t \\ 0 & \text{sinon.} \end{cases}$$

Soit $\widehat{X}$ la matrice dont les colonnes représentent les nouvelles variables Φ_r et les lignes les nouvelles coordonnées des objets symboliques :

$$\widehat{X} = [\Phi_1 | \ldots | \Phi_r | \ldots | \Phi_{K-p+1}].$$

On applique l'analyse factorielle discriminante classique à $\widehat{X}$.

Les solutions sont alors données par la résolution de :

$$T^{-1}BF = \lambda F$$

où :

$$T = \widehat{X}'H\widehat{X}$$

avec H la matrice des poids des différents sommets : $H = \frac{1}{N}\, I_N$

$$B = (\widehat{X}'HC)(C'HC)^{-1}(C'H\widehat{X}).$$

Nous devons donc trouver les valeurs λ_l et vecteurs propres v_l de :

$$(\widehat{X}'H\widehat{X})^{-1}(\widehat{X}'HC)(C'HC)^{-1}(C'H\widehat{X}).$$

Les coordonnées des sommets des hypercubes associés aux différents objets symboliques dans le plan factoriel à s dimensions sont alors données par :

$$y_i = V_s'(\widehat{X})_i \quad l = 1, \ldots, s$$

où V_s est la matrice contenant en colonne les vecteurs propres que nous avons retenus avec s inférieur à $m-1$. Le nombre de composantes principales que

l'on gardera ne pourra, en effet, pas dépasser $m-1$ puisque B a, au plus, $m-1$ valeurs propres non nulles.

Comme nous l'avons vu, les objets symboliques sont représentés sur le plan factoriel par le rectangle d'aire maximum couverte des sommets associés à chaque objet de E.

Pour représenter les classes sur le plan factoriel, on utilise également des rectangles d'aire maximum couverte. Les classes seront donc représentées par des rectangles contenant tous les individus appartenant à cette classe (représentés également par des rectangles).

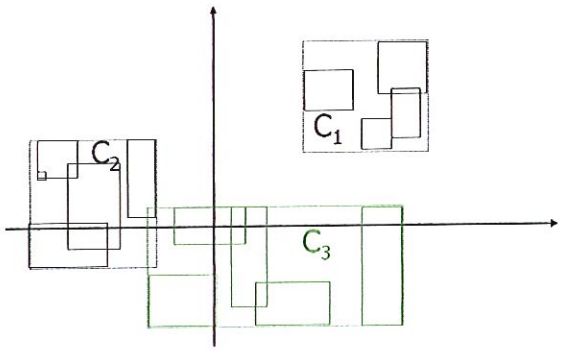


FIG. 4.1 – Représentation des classes par des rectangles d'aire maximum couverte (R. Verde, Analyse des données symboliques, Université de Namur, 2006)

4.3.6 Règle d'affectation

Nous avons à présent obtenu une représentation des objets symboliques dans un espace de dimension inférieure. Dans l'analyse factorielle discriminante symbolique, les objets appartenant à l'ensemble d'entraînement sont affectés à leurs classes d'appartenance connue a priori et nous allons chercher à classer de *nouveaux* individus sur lesquels les mêmes variables ont été mesurées.

Remarque :

Si nous désirons voir la classe à laquelle serait affectée un individu de l'ensemble d'apprentissage, il nous faut appliquer l'analyse factorielle à l'ensemble d'apprentissage sans cet élément et le reclasser.

Dans le même ordre d'idée que pour l'analyse factorielle discriminante clas-

sique, nous allons affecter un individu à la classe dont il est le plus proche. Mais ici, les classes et les objets étant représentés par des rectangles, nous ne pourrons pas utiliser une distance euclidienne ou une distance de Mahalanobis.

Nous classerons un objet v dans la classe Π_t si il est inclus dans une et une seule représentation d'un objet symbolique u de E qui appartient à la classe Π_t c'est-à-dire si l'objet v se trouve dans le rectangle d'aire maximum couverte de u.

Sinon, si l'image de v est en dehors de tous les rectangles d'aire maximum couverte de tous les objets symboliques de E ou si elle est sur une aire de recouvrement de 2 rectangles d'aire maximum couverte (ou plus) d'objets symboliques appartenant à des classes différentes, nous devons définir des mesures de similarité entre les éléments et les classes : on assignera alors l'objet v appartenant à la classe Π_i dont les éléments ont la plus grande similarité par rapport à v.

Descripteur de potentiel π

Pour définir une règle de classification géométrique, nous proposerons deux approches basées sur le concept de *descripteur de potentiel* $\pi(.)$ défini comme le volume du produit cartésien des domaines des variables. Ce potentiel peut être défini pour un objet u sur les axes factoriels par :

$$\pi(u) = \prod_{l=1}^{s} \mu(S_{ul})$$

où $S_{u\alpha}$ est le domaine du l^e axe factoriel couvert par l'individu u et $\mu(S_{ul})$ est une mesure de S_{ul} définie par :

$$\mu(S_{ul}) = \begin{cases} \text{au nombre de catégories} & \text{si nous avons des variables nominales} \\ & \text{ou multivaluées} \\ \text{la longueur de l'intervalle} & \text{pour des variables intervalles.} \end{cases}$$

Définition de 2 règles de classification

Une règle de classification basée sur une extension de la mesure de dissimilarité de Minkowsky

Cette mesure est généralement appelée mesure de dissimilarité d'Ichino-Carvalho.

La dissimilarité d'un objet u de l'ensemble E à un objet v appartenant aux différentes classes :

$$d(v,u) = \sqrt[r]{\sum_{l=1}^{s} \left(p_l \phi(S_{vl}, S_{ul})\right)^r}$$

où :

- $p_l > 0$ sont des poids $(l = 1, \ldots, s)$
- $r \geq 1$
- ϕ est une fonction définie par :

$$\phi(S_{vl}, S_{ul}) = \frac{\mu(S_{vl} \oplus S_{ul}) - \mu((S_{vl} \cap S_{ul}) + \gamma(2\mu(S_{vl} \cap S_{ul} - \mu(S_{vl}) - \mu(S_{ul}))}{\mu(S_{vl} \oplus S_{ul})}$$

avec $\gamma \in [0,1]$ un poids.

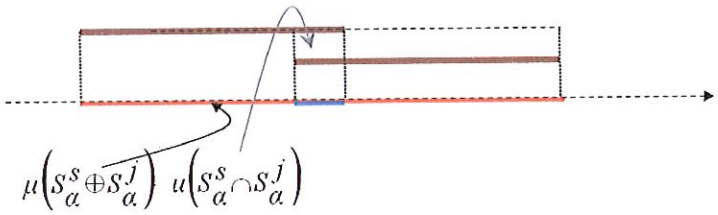


FIG. 4.2 – Ilustration de la fonction Φ (R. Verde, Analyse des données symboliques, Université de Namur, 2006)

Pour prendre en compte la différence de variance des s axes factoriels, nous prendrons les valeurs propres λ_l obtenues dans l'analyse factorielle classique comme poids p_l $(l = 1, \ldots, s)$.
Nous choisirons une valeur pour γ suivant l'importance que nous voulons donner au terme $\mu(S_{vl} \cap S_{ul}) - \mu(S_{vl}) - \mu(S_{ul})$.

Les mesures de dissimilarité précédentes sont calculées pour toutes les images des objets symboliques u appartenant à la classe C_t.

Rappel :

$\oplus$ désigne une somme directe. Nous aurons que S_{vl} et S_{ul} sont en somme directe si et seulement si pour tout élément w de $S_{vl} + S_{ul}$ il existe un unique couple (u_1, v_1) de $S_{vl} \times S_{ul}$ tel que $w = u_1 + v_1$. c'est-à-dire si la décomposition de tout élément de $S_{vl} + S_{ul}$ en un élément de S_{vl} et en un élément de S_{ul} est unique.

On calcule alors la distance moyenne de v à tous les éléments de la classe C_t :

$$\overline{d}(v, C_t) = \frac{1}{n_t} \sum_{u \in C_t} d(v, u) \quad \text{avec } n_t = |C_t|.$$

Finalement, nous assignerons v à la classe Π_t pour laquelle la dissimilarité moyenne $d(v, C_t)$ est minimale.

Une règle de classification basée sur l'augmentation minimale du descripteur de potentiel

Une autre règle est basée sur l'augmentation du descripteur de potentiel (pdi) des images des objets symboliques $u \in C_t$ sur le sous-espace factoriel lorsqu'on ajoute l'image de l'élément v que l'on souhaite classer.

Cette mesure est définie par :

$$pdi(S_u, S_v) = \frac{\prod_{l=1}^{s} \mu(S_{ul} \oplus S_{vl}) - \prod_{l=1}^{s} \mu(S_{vl})}{\prod_{l=1}^{s} \mu(S_{ul})}$$

et doit être calculée pour tous les éléments $u \in E$.

Elle représente, par exemple pour $\alpha = 2$, l'aire du rectangle qui possède l'extension maximale des objets symboliques u de la classe C_t qui est inclus dans le rectangle associé avec l'objet v sur le plan factoriel.

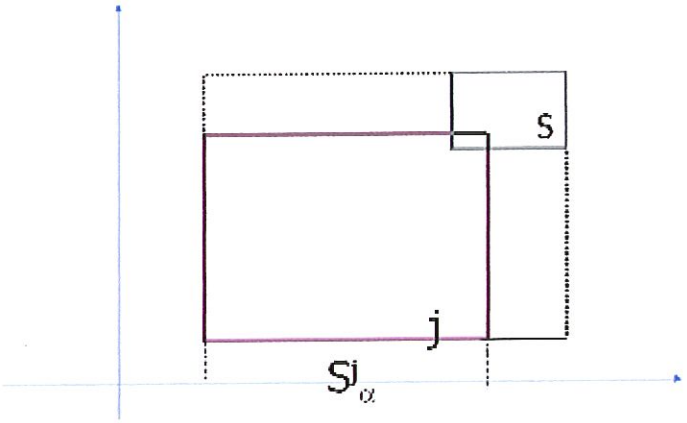


FIG. 4.3 – Illustration de du descripteur de potentiel ((R. Verde, Analyse des données symboliques, Université de Namur, 2006))

Remarques :

- Si $\prod_{l=1}^{s} \mu(S_{ul} \oplus S_{vl}) = \prod_{l=1}^{s} \mu(S_{ul})$ alors S_v est contenu dans l'ensemble d'extension de l'image de u.

- Si $pdi(S_u, S_v) = 0$ alors l'image de v coïncide parfaitement avec l'image de u.

Un objet symbolique v sera assigné à la classe de l'objet symbolique u pour laquelle $pdi(S_u, S_v)$ est minimal pour tous les éléments $u \in E$.

4.3.7 La validation

La dernière étape de notre analyse consiste alors à évaluer la qualité des résultats que nous venons d'obtenir. Pour ce faire, nous utiliserons les mêmes méthodes que pour l'analyse factorielle discriminante classique, par exemple en définissant un ensemble test ou en utilisant la méthode du leave-one-out. Ces méthodes restent applicables puisque nous avons considéré ici le cas où les différentes classes C_t sont composées d'objets symboliques u du premier ordre.

4.3.8 Lien avec la méthode classique

Comme nous l'avons fait pour l'analyse en composantes principales classique, nous pourrions montrer que l'analyse factorielle discriminante classique est un cas particulier de la méthode symbolique. Nous ne détaillerons pas la preuve puisqu'elle nécessiterait une connaissance plus approfondie de l'analyse canonique généralisée symbolique pour l'étape de la quantification des variables.

Nous pouvons cependant remarquer que l'analyse factorielle classique pour des variables quantitatives correspond à une analyse factorielle symbolique où toutes les variables sont des variables intervalles où la borne inférieure est égale à la borne supérieure. Pour la preuve, nous devrions montrer que nous appliquons l'analyse factorielle classique à la matrice de données classiques X. Autrement dit, nous devrions montrer que l'étape de quantification des variables nous permet de retrouver la matrice X ($X = \widehat{X}$).

4.4 Exemples

4.4.1 Sur des données artificielles

Reprenons l'exemple de nos 5 films u, v, w, x et z, sur lesquels nous avions mesurés 4 variables :

- Y_1 une variable qualitative représentant le genre d'un film
- Y_2 une variable modale mesurant la probabilité de choisir un film dans une des 7 catégories de Y_1
- Y_3 une variable quantitative donnant le bénéfice moyen des locations des 5 films
- Y_4 une variable intervalle décrivant le nombre de locations par semaine des 5 films.

Les 5 films sont répartis en 3 classes formées de la manière suivante :

- C_1 regroupe les films familiaux et les comédies
- C_2 les films d'action de science-fiction et les thrillers
- C_3 les films catastrophes et les drames.

Nous avons donc que x et z appartiennent à C_1, v et w à C_2 et u à C_3.

Rappelons brièvement les différentes étapes de l'analyse factorielle discriminante symbolique :

1. On code toutes les variables et on construit la matrice de codage globale X.
2. On quantifie les variables en calculant les vecteurs propres Φ_r de

$$\frac{1}{N} G'X\Delta_x^I X'G$$

ce qui nous permet de construire la matrice $\widehat{X}$.
3. On applique l'analyse factorielle classique. Pour cela, on calcule les valeurs propres et les vecteurs propres de :

$$(\widehat{X}'H\widehat{X})^{-1}(\widehat{X}'HC)(C'HC)^{-1}(C'H\widehat{X}).$$

Les l^e coordonnées des sommets dans le plan factoriel seront alors donnés par :

$$\Psi_l = \widehat{X} v_l.$$

4. On définit notre règle d'affectation afin de classer nos nouveaux individus.
5. On valide les résultats que l'on a obtenu soit par un échantillon test soit par la méthode du leave-one-out ou encore par la méthode du bootstrap.

Première étape : Codage des variables symboliques

Nous avions obtenu la matrice de codage X suivante :

$$\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 3/4 & 1/4 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 5/8 & 3/8 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 3/4 & 1/4 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 5/8 & 3/8 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0.6 & 0.4 & 0 & 5/8 & 3/8 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0.6 & 0.4 & 0 & 0 & 7/8 & 1/8 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 3/4 & 1/4 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 0 & 3/4 & 1/4 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 3/4 & 1/4 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 0 & 3/4 & 1/4 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/3 & 2/3 & 0 & 0 & 7/8 & 1/8 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/3 & 2/3 & 0 & 0 & 3/4 & 1/4 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 3/4 & 1/4 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 1 & 0 & 1/2 & 1/2 & 0
\end{pmatrix}$$

Chaque ligne correspondant aux coordonnées des 14 sommets représentent les 5 objets symboliques dans un espace de dimension 20.

Nous avions aussi obtenu la matrice G :

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Deuxième étape : Quantification des variables symboliques

Nous devons trouver les valeurs propres et les vecteurs propres de :

$$\frac{1}{N} G' X \Delta_x^I X' G$$

où Δ_x^I est une matrice diagonale par blocs de dimension 20×20 où les blocs sont $(X_j' X_j)^1 \quad j = 1, \ldots, 4$:

$$X_1' X_1 = I_7 \quad X_2' X_2 = 49 I_7$$

$$X_3' X_3 = \begin{pmatrix} 2.067 & -4.55 & 0.6825 \\ -0.455 & 0.4841 & -0.7262 \\ 0.6825 & -0.7262 & 7.3393 \end{pmatrix}$$

$$X_4' X_4 = \begin{pmatrix} 0.7991 & -0.9152 & 3.4777 \\ -0.9152 & 1.942 & -7.3795 \\ 3.4777 & -7.3795 & 34.442 \end{pmatrix}$$

Nous devons donc trouver les valeurs propres λ_r et les vecteurs propres Phi_r $(r = 1, \ldots, 5)$ de :

$$\begin{pmatrix} 9.3237 & 3.9967 & 8.6655 & 4.0531 & 4.0893 \\ 3.9967 & 2.5075 & 3.9674 & 2.1708 & 2.0312 \\ 8.6655 & 3.9674 & 9.3492 & 4.1233 & 4.0357 \\ 4.0531 & 2.1708 & 4.1233 & 2.4780 & 2.0045 \\ 4.0893 & 2.0312 & 4.0357 & 2.0045 & 2.6161 \end{pmatrix}.$$

Ce qui nous permet de construire les variables quantifiées par :

$$\Phi_r = X\ \Delta_x^{-1}\ X'\ G\ \Phi_r \quad \text{avec} \quad r = 1, \ldots, 5.$$

On obtient alors la matrice $\widehat{X}$ suivante où la j^{e} colonne représente la j^{e} variable quantifiée :

$$\widehat{X} = \begin{pmatrix} -38.3279 & -2.0286 & -12.3037 & 19.9100 & -23.9988 \\ -38.0784 & -1.9069 & -12.4679 & 19.9562 & -23.8526 \\ -38.3279 & -2.0286 & -12.3037 & 19.9100 & -23.9988 \\ -38.0784 & -1.90698 & -12.4679 & 19.9562 & -23.8526 \\ -34.8343 & 1.3829 & -12.7699 & 22.2734 & -20.9766 \\ -34.1693 & 1.9589 & -12.7537 & 22.4742 & -20.9065 \\ -38.3838 & -1.2391 & -9.8412 & 21.0494 & -24.1075 \\ -38.0516 & -0.5737 & -9.1518 & 21.2664 & -24.5524 \\ -38.3838 & -1.2391 & -9.8412 & 21.0494 & -24.1075 \\ -38.0516 & -0.5737 & -9.1518 & 21.2664 & -24.5524 \\ -33.6846 & 1.1097 & -12.3972 & 22.0392 & -23.8068 \\ -34.2669 & 1.0773 & -11.5597 & 22.0093 & -24.4680 \\ -34.6040 & -3.6610 & -11.400 & 23.5198 & -21.6957 \\ -34.1051 & -3.4176 & -11.9685 & 23.6120 & -21.4032 \end{pmatrix}$$

Troisième étape : Application de l'analyse factorielle discriminante classique

Pour cela, nous devons tout d'abord construire la matrice C indiquant l'appartenance des différents sommets des films aux différentes classes C_i. Nous aurons donc :

$$C_{it} = \begin{cases} 0 & \text{si le film auquel appartient le sommet } i \text{ appartient à la classe } C_t \\ 0 & \text{sinon.} \end{cases}$$

Nous obtenons alors :

$$C = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

La matrice diagonale des poids H est donnée par : $\frac{1}{14} I_{14}$. Nous pouvons alors appliquer l'analyse factorielle discriminante classique en calculant :

- la matrice $T^{-1}B$ ainsi que ses valeurs et vecteurs propres λ_l et v_l
- le nombre d'axes factoriels nécessaire (ici, 2 axes nous permettront d'expliquer environ 99% de la variance totale)
- les coordonnées des sommets dans un espace de dimension 2 par :

$$y_l = \widehat{X}\, v_l \; ; \;\; l = 1, 2.$$

Nous obtenons alors les résultats suivants :

	F_1	F_2
y_{u1}	-0.5489	-0.0913
y_{u2}	-0.5816	-0.0913
y_{u3}	-0.5489	-0.0913
y_{u4}	-0.5816	-0.0913
y_{v1}	0.2803	-0.4385
y_{v2}	0.2902	-0.4385

y_{w1}	0.1822	-0.4385
y_{w2}	0.3326	-0.4385
y_{w3}	0.1822	-0.4385
y_{w4}	0.3326	-0.4385
y_{x1}	0.0678	0.7490
y_{x2}	0.2410	0.7490
y_{z1}	0.2088	0.7490
y_{z2}	0.1434	0.7490

Nous pouvons alors représenter nos films par des rectangles. Ici, ces rectangles correspondront à des droites puisque nous n'avons que deux sommets différents par film. Cela vient du fait que les coordonnées des sommets sont déjà très proches dans la matrice de codage global X.

Nous pouvons alors former nos classes, de nouveau par des rectangles et ces rectangles correspondent également à des droites.

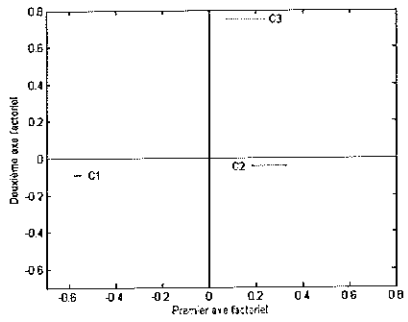


FIG. 4.4 – Représentation des différentes classes par des rectangles

Quatrième étape : Définition d'une règle d'affectation

Pour cela, supposons que nous ayons 3 nouveaux films que nous noterons t_1, t_2 et t_3. t_1 est une comédie, t_2 est un film d'action et t_3 un drame. A priori, ces films appartiennent aux classes C_1, C_2 et C_3 respectivement.

Les coordonnées des sommets les représentant sur le plan factoriel sont données par :

y_{t11}	-0.3022	-0.2024
y_{t12}	-0.2714	-0.1814
y_{t21}	0.1814	0.4178
y_{t22}	0.3221	0.4312
y_{t31}	-0.2314	0.5114
y_{t32}	-0.1708	0.6724

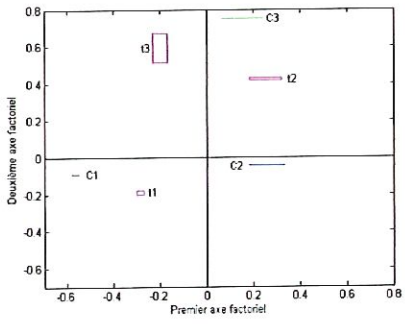


FIG. 4.5 – Représentation des nouveaux films sur le plan factoriel

Pour classer ces différents films, nous utilisons la règle de classification basée sur l'augmentation minimale du descripteur de potentiel. Commençons par évaluer les différentes mesures S_{kj} avec $k = u, v, w, x, z, t_1, t_2, t_3$ et $j = 1, 2$:

$$S_{u1} = 0.4576 \qquad S_{u2} = 0.000001$$
$$S_{v1} = 0.0099 \qquad S_{v2} = 0.000001$$
$$S_{w1} = 0.1504 \qquad S_{w2} = 0.000001$$
$$S_{x1} = 0.1732 \qquad S_{x2} = 0.000001$$
$$S_{z1} = 0.0654 \qquad S_{z2} = 0.000001$$
$$S_{t11} = 0.0308 \qquad S_{t12} = 0.0210$$
$$S_{t21} = 0.1407 \qquad S_{t22} = 0.0134$$
$$S_{t31} = 0.0134 \qquad S_{t32} = 0.1610$$

Remarque :
Les mesures sur le deuxième axe factoriel sont posées à 0.000001 puisque les objets sont représentés par des droites.

Nous pouvons alors calculer les descripteurs de potentiel $pdi(k, k')$ où $k = u, v, w, x, y$ et $k' = t_1, t_2, t_3$.

Pour $pdi(u, t_1)$, nous avons :

$$pdi(u, t_1) = \frac{\prod_{i=1}^{2} \mu(S_{ui} + S_{t1i}) - \mu(S_{t_1 i})}{\mu(S_{ui})}$$

où $\mu(S_{ui} + S_{t1i})$ est la longueur du domaine couvert sur le i^e axe factoriel entre le film u et le film t_1. Ce qui nous donne :

$$pdi(u, t_1) = \frac{(0.4108)(0.7637)}{9,7566.10^{-3}} = 5848,2452$$

Nous obtenons alors :

	t_1	t_2	t_3
u	5848,2452	1027748,842	664273,0769
v	1531911,5150	3672066,667	57544327,27
w	910250,6649	796447,141	4101003,989
x	2980102,0790	475397,1132	591718,4758
z	742383,1804	1522090,826	1450075,229

Nous avons donc que t_1 est affecté à la classe de u c'est-à-dire C_1, t_2 et t_3 à celle de x, C_3.

Ces résultats sont bien logiques puisque, comme nous pouvons le voir pour t_1, l'augmentation de volume de la classe C_3 est minimale lorsque l'on ajoute t_1.

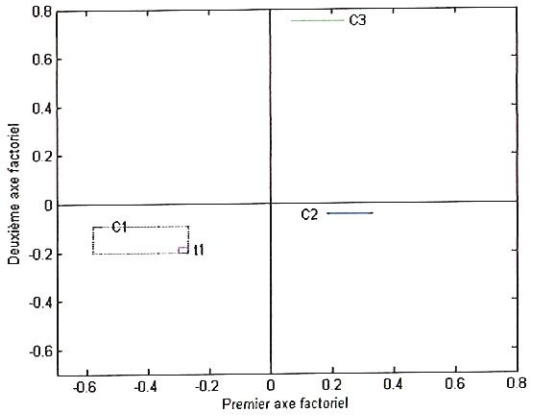


FIG. 4.6 – Représentation de la classe C_1 lorsque l'on y ajoute le film t_1

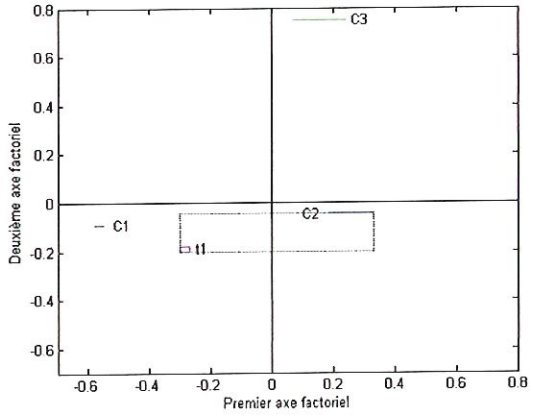


FIG. 4.7 – Représentation de la classe C_2 lorsque l'on y ajoute le film t_1

Cinquième étape : la validation

Nous pouvons utiliser la méthode de l'ensemble test à partir des films t_1, t_2 et t_3 et nous obtenons alors un taux de mauvais classement de 0.33 puisque seul t_2 n'est pas affecté à la bonne classe.

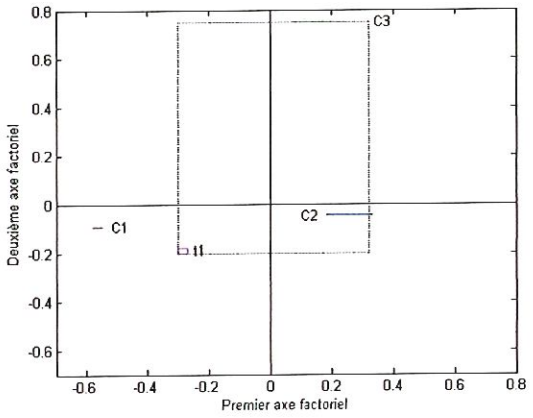


FIG. 4.8 – Représentation de la classe C_3 lorsque l'on y ajoute le film t_1

4.4.2 Sur des données réelles : Caractéristiques de voitures

Analysons un ensemble de 23 voitures sur lesquelles nous avons mesuré les variables suivantes :

- Y_1 le prix
- Y_2 la cylindrée
- Y_3 le carburant
- Y_4 la traction
- Y_5 la vitesse maximale
- Y_6 l'accéleration
- Y_7 l'empatement
- Y_8 la longueur
- Y_9 la largueur
- Y_{10} la hauteur
- Y_{11} la catégorie

Les variables $Y_1, Y_2, Y_5, Y_6, Y_7, Y_8, Y_9$ et Y_{10} sont des variables intervalles, Y_3 et Y_4 sont des variables catégoriques dont les différentes catégories sont les suivantes :

Pour Y_3 :
- Essence
- Diesel

Pour Y_4 :
- traction avant

- traction arrière
- traction avant et arrière

Y_{11} est une variable qualitative dont les catégories sont :

- Utilitaire
- Berline
- Limousine
- Sportive

L'appartenance aux classes a priori est donnée en fonction de la variable Y_{11} c'est-à-dire de la catégorie de la voiture. Nous aurons donc quatre classes :

- les utilitaires
- les berlines
- les limousines
- les sportives

Ces données sont extraites d'une base de donnée contenue dans les bases du logiciel Sodas sous le nom `cars.sds`.

La première chose que nous devons faire est le codage des variables. Les variables intervalles seront codées à l'aide de fonctions B-splines et chacune des matrices de codage correspondante sera de dimension 66×3.
Pour les variables catégoriques et qualitatives, nous utiliserons un codage binaire.

Une fois ces matrices codées, nous pouvons construire notre matrice de codage globale X, cette matrice sera de dimension $N \times K$ où :

$$N = \sum_{i=1}^{33} \left(2^8 + \sum_{j=1}^{3} k_{ij}\right)$$

où k_j est le nombre de catégorie de la variable Y_j intervenant pour la voiture i et

$$K = 33.$$

La deuxième étape consiste à quantifier nos variables à l'aide de $K - p + 1$ vecteurs propres de dimension $n \times 1$ pour pouvoir appliquer l'analyse factorielle discriminante classique. Nous aurons donc ici 23 vecteurs propres de dimensions 33×1.

Nous devons ensuite appliquer l'analyse factorielle discriminante classique à ces variables quantifiées et nous avons que deux axes factoriels nous permettent d'expliquer environ 95% de la variance totale des données.

Nous pouvons alors calculer les coordonnées des sommets sur le plan factoriel et représenter les 33 voitures par des rectangles.

Et nous pouvons définir nos classes C_i par des rectangles sur le plan factoriel :

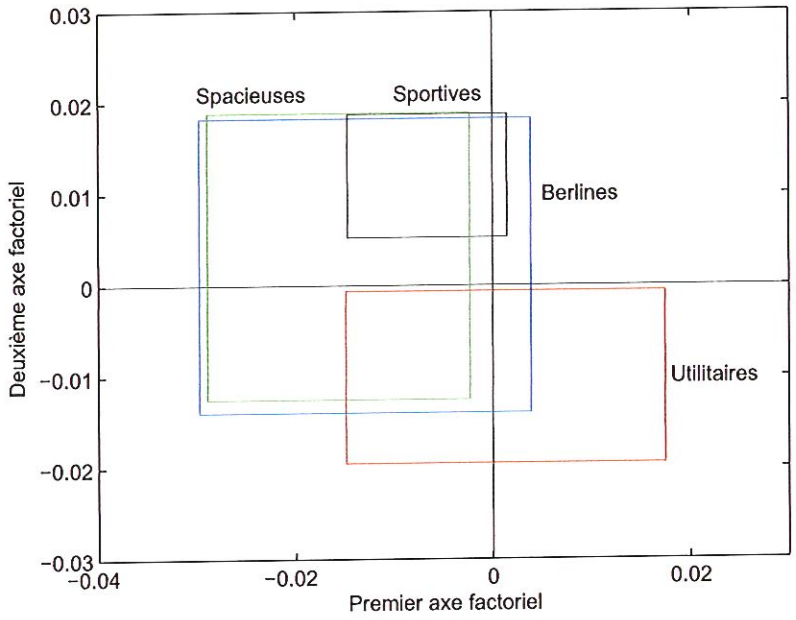


FIG. 4.9 – Représentation des 4 classes

Si nous construisons une règle d'affectation à partir de la mesure de dissimilarité d'Ichino-Carvalho, nous obtenons un taux de mauvais classements d'environ 30%.

Les différentes classes sont formées des voitures suivantes :

C_1 : Alpha 156, Fiesta, Punto, Lancia Yaris, Nissan Micra, Corsa, Porsche, Twingo, Rover 25, Skoda Fabia, Skoda octavia
C_2 : Alpha 145, Audi A3, Audi A8, BMW série 3, Focus, Punto, Opel Vectra, Rover 75, Mercedes SL et Passat
C_3 : Alfa 166, Mercedes classe C, E et S, Lancia K, Audi A6, BMW serie 5 et 7
C_4 : Aston Martin, Lamborghini, Ferrari, Maserati G et Honda NSK.

Une première remarque à faire est qu'aucune des voitures ne se retrouve classée en sportive (C_4) si elle ne l'était pas a priori.
Les seules voitures qui sont classées a posteriori dans la classe des limousines (C_3) et qui ne l'étaient pas a priori sont des berlines (C_2).
Les mauvais classements proviennent surtout des deux permières classes. Nous avons en effet la matrice de classification suivante :

	C_1	C_2	C_3	C_4
C_1	8	2	0	0
C_2	2	4	2	0
C_3	0	2	6	0
C_4	1	1	0	5

où les lignes correspondent aux classements a priori et les colonnes aux classements a posteriori.

Ces mauvais classements sont quelque part assez logiques. En effet, si certaines limousines sont plus courtes ou si des voitures utilitaires sont plus grandes alors elles peuvent se retrouver classées en berlines.

Troisième partie

Applications

Chapitre 5

Utilisation du logiciel Sodas

5.1 Introduction

Sodas 2 est un logiciel prototype public issu du projet ASSO (Analysis System of Symbolic Official data). Ce projet est la continuation d'un premier projet Eurostat appelé Sodas.

Le logiciel Sodas 2 peut être téléchargé gratuitement à l'adresse suivante :

www.info.fundp.ac.be/asso/

Sodas a pour but l'analyse des données symboliques. Il travaille donc à partir d'un fichier de données symboliques (des fichiers .sds). Pour obtenir ces fichiers, nous devons, par exemple, construire une base de données à l'aide de Microsoft access et utiliser le logiciel DB2SO inclus dans Sodas. Les différentes étapes de la construction de tels fichiers seront dévelopées en annexe.

Remarque :

N'oublions pas que les données classiques sont un cas particulier des données symboliques. Le logiciel fonctionne donc également avec des données classiques lorsqu'elles sont représentées sous forme symbolique.

A partir du fichier de données, nous pouvons appliquer toute une série de méthodes d'analyse de données comme la classification, la régression linéaire, la construction de différents type de graphiques, ...

Sodas suit donc le schéma suivant :

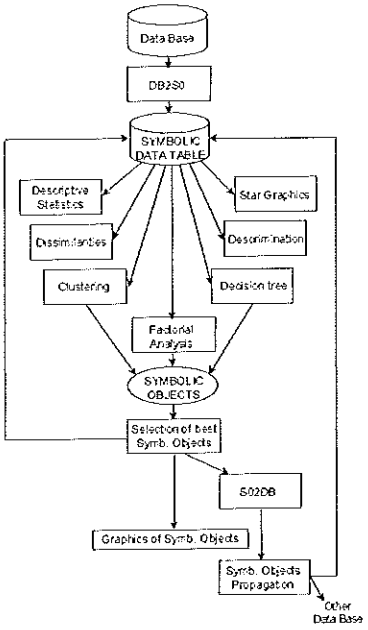


FIG. 5.1 – Etapes effectuées par le logiciel Sodas

Principes de base du logiciel

Dans Sodas, une analyse de données se représente par une filière. Les différentes méthodes possibles sont représentées par des icônes et ces icônes sont liées dans la filière (`chaining`). Les différentes méthodes possibles sont reprises dans le menu déroulant de la fenêtre `methods`. Elles sont répertoriées suivant leur type (`descriptive statistics`, `clustering`, `factorial`, ...)

Pour sélectionner la méthode, il suffit de faire glisser le carré correspondant à la méthode choisie dans l'onglet `chaining` entre la base et le `end`.

La première chose à faire pour appliquer une méthode est d'ouvrir une base de données. Pour cela, il suffit de faire un clic droit sur l'icône base et de rechercher cette base.

Elle apparaît alors dans l'onglet `chaining`.

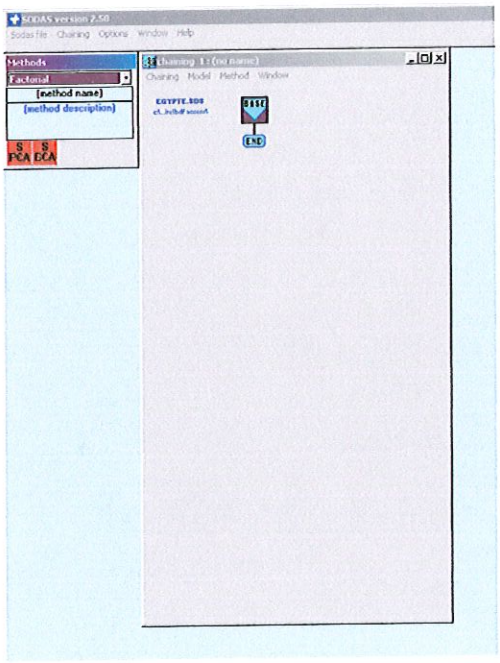


FIG. 5.2 – Sélection d'une base

Nous pouvons ensuite choisir la (ou les) méthode(s) que nous souhaitons appliquer. Ces méthodes seront executées suivant l'ordre dans lequel elles apparaissent de haut en bas.

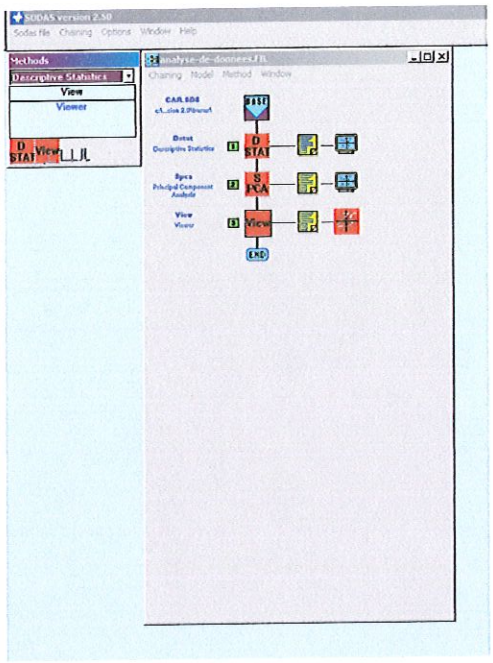


FIG. 5.3 – Exemple de filière executée par Sodas

5.2 Visualisation des objets symboliques

Pour visualiser les différentes valeurs des variables contenues dans notre base de données, Sodas propose la méthode `view`. Cette méthode nous permet d'obtenir le tableau de données symboliques représentant notre base ainsi que des graphiques appellés zoom stars.

Ces graphiques en forme d'étoile sont une représentation des valeurs des différentes variables pour chaque individu (1 variable correspond à une branche de l'étoile).

Nous pouvons obtenir ces graphiques en étoiles :

- en deux dimensions : les valeurs des différentes variables sont alors reliées entre elles
- ou en trois dimensions : les valeurs des variables sont alors représentées sous forme d'histogramme.

La méthode `view` se situe dans la section `Descriptive statistics` des méthodes.

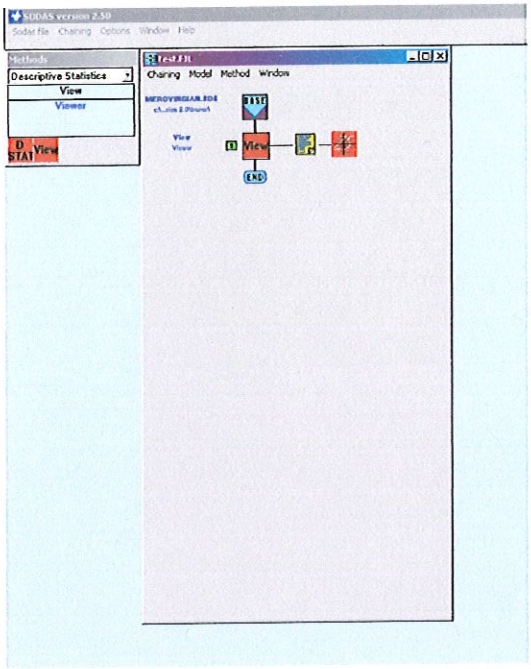


FIG. 5.4 – La méthode `view`

Nous devons tout d'abord fixer les paramètres de la méthode. Pour cela, il suffit de faire un clic droit sur l'icône de la méthode et de choisir `parametres`. Une fois ces paramètres fixés, la méthode apparaît en rouge dans notre chaîne.
Les paramètres à choisir ici sont les individus et les variables que nous souhaitons représenter. Nous obtenons la liste des variables présentes dans notre base de données (`avalaible variables`) et celles que nous choisissons (`selected variables`).

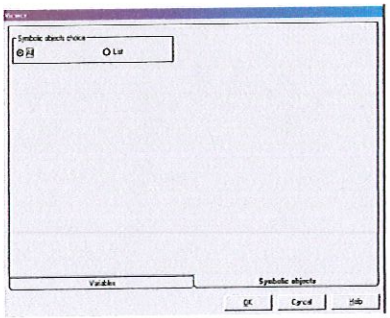


FIG. 5.5 – Choix des objets

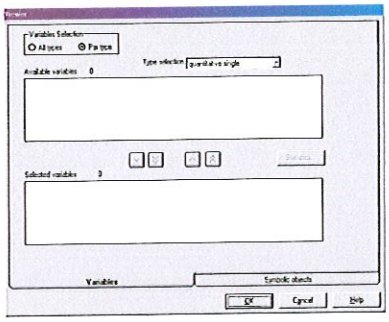
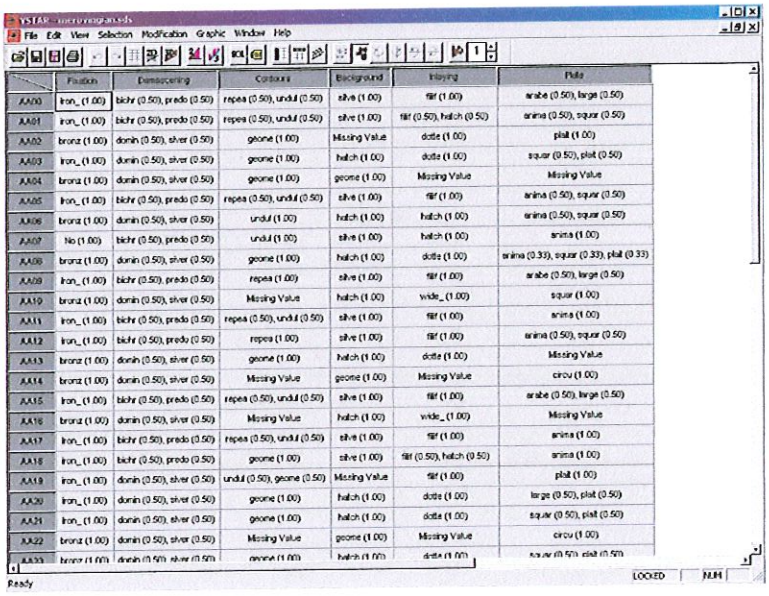


FIG. 5.6 – Choix des variables

Nous pouvons alors exécuter la méthode. Pour cela, il suffit d'aller dans l'onglet `Method` et de sélectionner `run chaining`. Nous devrons alors sauver la filière (en `.fil`) et, éventuellement, lui donner un titre.

Après avoir exécuté la méthode, nous obtenons deux sorties : une sortie texte et une sortie graphique. La sortie texte nous donne simplement le nombre de variables et d'individus que nous avions sélectionnés. La sortie graphique nous donne le tableau de nos données symboliques.



	Position	Osmoscoring	Contours	Background	Inlaying	Plate
AA00	iron_ (1.00)	bichr (0.50), predo (0.50)	repea (0.50), undul (0.50)	silve (1.00)	filif (1.00)	arabe (0.50), large (0.50)
AA01	iron_ (1.00)	bichr (0.50), predo (0.50)	repea (0.50), undul (0.50)	silve (1.00)	filif (0.50), hatch (0.50)	anima (0.50), squar (0.50)
AA02	bronz (1.00)	domin (0.50), silver (0.50)	geome (1.00)	Missing Value	dotte (1.00)	plait (1.00)
AA03	iron_ (1.00)	domin (0.50), silver (0.50)	geome (1.00)	hatch (1.00)	dotte (1.00)	squar (0.50), plait (0.50)
AA04	bronz (1.00)	domin (0.50), silver (0.50)	geome (1.00)	geome (1.00)	Missing Value	Missing Value
AA05	iron_ (1.00)	bichr (0.50), predo (0.50)	repea (0.50), undul (0.50)	silve (1.00)	filif (1.00)	anima (0.50), squar (0.50)
AA06	bronz (1.00)	domin (0.50), silver (0.50)	undul (1.00)	hatch (1.00)	hatch (1.00)	anima (0.50), squar (0.50)
AA07	No (1.00)	bichr (0.50), predo (0.50)	undul (1.00)	silve (1.00)	hatch (1.00)	anima (1.00)
AA08	bronz (1.00)	domin (0.50), silver (0.50)	geome (1.00)	hatch (1.00)	dotte (1.00)	anima (0.33), squar (0.33), plait (0.33)
AA09	iron_ (1.00)	bichr (0.50), predo (0.50)	repea (1.00)	silve (1.00)	filif (1.00)	arabe (0.50), large (0.50)
AA10	bronz (1.00)	domin (0.50), silver (0.50)	Missing Value	hatch (1.00)	wide_ (1.00)	squar (1.00)
AA11	iron_ (1.00)	bichr (0.50), predo (0.50)	repea (0.50), undul (0.50)	silve (1.00)	filif (1.00)	anima (1.00)
AA12	iron_ (1.00)	bichr (0.50), predo (0.50)	repea (1.00)	silve (1.00)	filif (1.00)	anima (0.50), squar (0.50)
AA13	bronz (1.00)	domin (0.50), silver (0.50)	geome (1.00)	hatch (1.00)	dotte (1.00)	Missing Value
AA14	bronz (1.00)	domin (0.50), silver (0.50)	Missing Value	geome (1.00)	Missing Value	circu (1.00)
AA15	iron_ (1.00)	bichr (0.50), predo (0.50)	repea (0.50), undul (0.50)	silve (1.00)	filif (1.00)	arabe (0.50), large (0.50)
AA16	bronz (1.00)	domin (0.50), silver (0.50)	Missing Value	hatch (1.00)	wide_ (1.00)	Missing Value
AA17	iron_ (1.00)	bichr (0.50), predo (0.50)	repea (0.50), undul (0.50)	silve (1.00)	filif (1.00)	anima (1.00)
AA18	iron_ (1.00)	bichr (0.50), predo (0.50)	geome (1.00)	silve (1.00)	filif (0.50), hatch (0.50)	anima (1.00)
AA19	iron_ (1.00)	domin (0.50), silver (0.50)	undul (0.50), geome (0.50)	Missing Value	filif (1.00)	plait (1.00)
AA20	iron_ (1.00)	domin (0.50), silver (0.50)	geome (1.00)	hatch (1.00)	dotte (1.00)	large (0.50), plait (0.50)
AA21	iron_ (1.00)	domin (0.50), silver (0.50)	geome (1.00)	hatch (1.00)	dotte (1.00)	squar (0.50), plait (0.50)
AA22	bronz (1.00)	domin (0.50), silver (0.50)	Missing Value	geome (1.00)	Missing Value	circu (1.00)

FIG. 5.7 – Tableau de données symboliques

A partir de ce tableau, nous pouvons également obtenir une représentation graphique des données en deux ou trois dimensions sous forme de zoom stars :

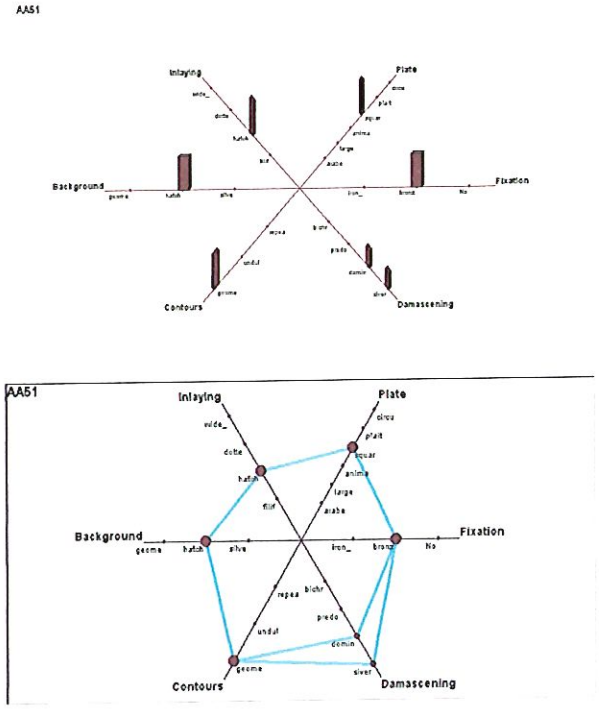


FIG. 5.8 – Exemples de zoom stars à deux ou trois dimensions

Pour cela, nous utilisons les icônes suivantes :

FIG. 5.9 – Représentation graphique des données symboliques

Nous obtenons alors un graphique par individu.

Nous pouvons également représenter tous les individus sur un même graphique.

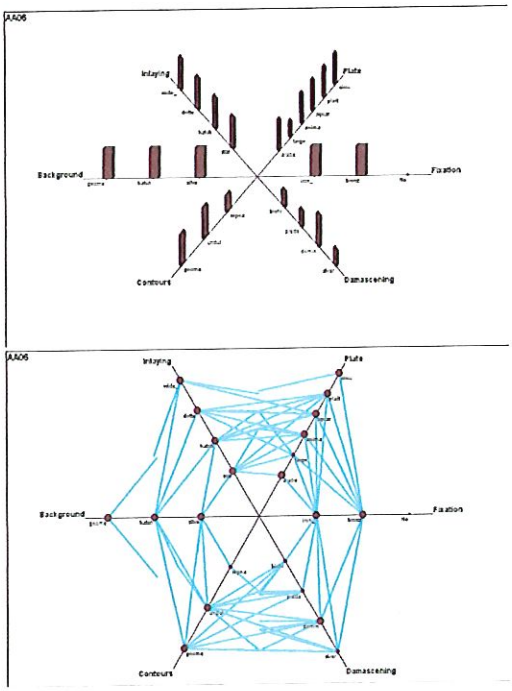


FIG. 5.10 – Exemples de zoom stars superposés à deux ou trois dimensions

Pour cela, nous utilisons :

FIG. 5.11 – Représentation graphique des données symboliques

Si nous ne souhaitons pas représenter tous les individus ou toutes les variables, nous pouvons les sélectionner à partir du tableau. Les variables et/ou individus sélectionnés apparaissent alors en gris.

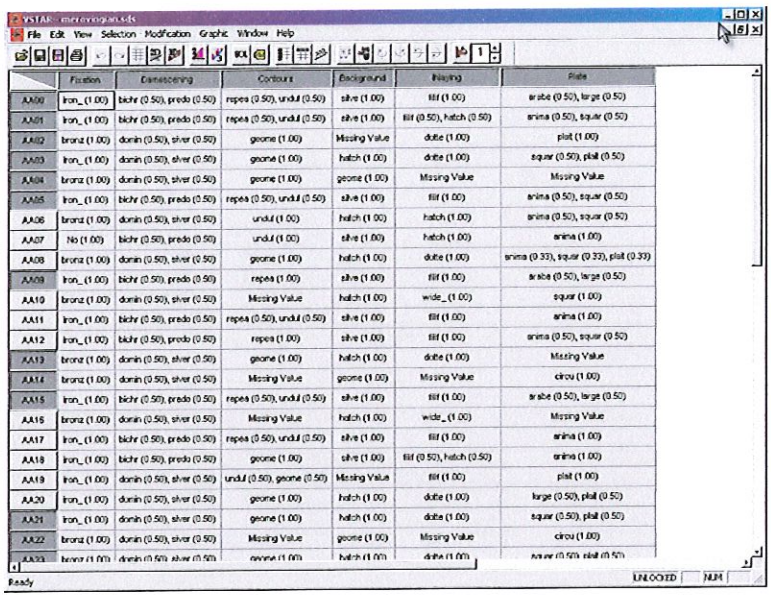


FIG. 5.12 – Sélection de certains individus et variables

De nombreuses autres options sont disponibles. Elles sont décrites dans le manuel utilisateur du logiciel Sodas.

5.3 L'analyse en composantes principales sous Sodas

Après avoir sélectionné notre base, nous devons sélectionner la méthode que nous souhaitons utiliser, ici l'analyse en composantes principales. Pour cela, nous choisissons `factorial` dans le menu déroulant de la fenêtre `methods`. Nous avons deux choix :

- `SPCA` : l'analyse en composantes principales symbolique
- et `SGCA` : l'analyse canonique généralisée symbolique.

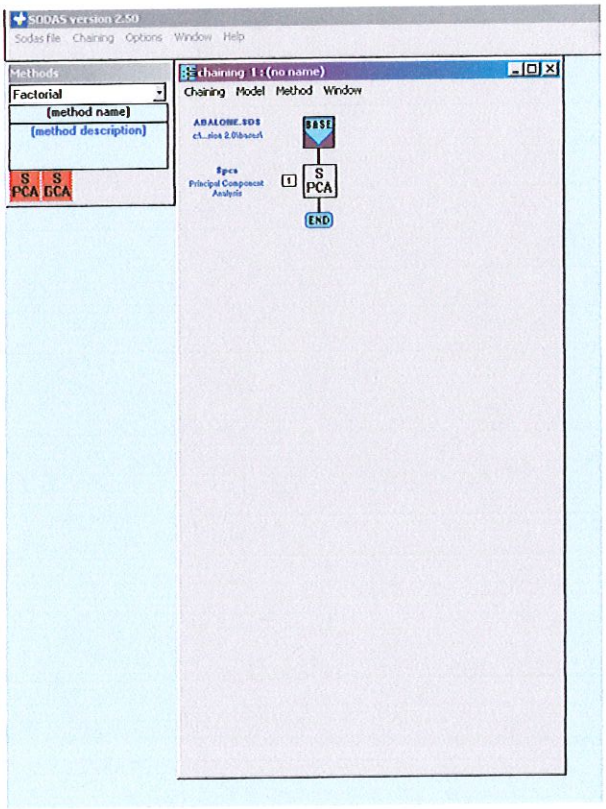


FIG. 5.13 – Sélection d'une méthode

Nous pouvons alors choisir les paramètres de l'analyse en composantes principales. Pour cela, il suffit de faire un clic droit sur l'icône de la méthode et de sélectionner `parametres`.

Le premier choix à faire est celui des variables intervalles qui interviendront dans notre analyse.

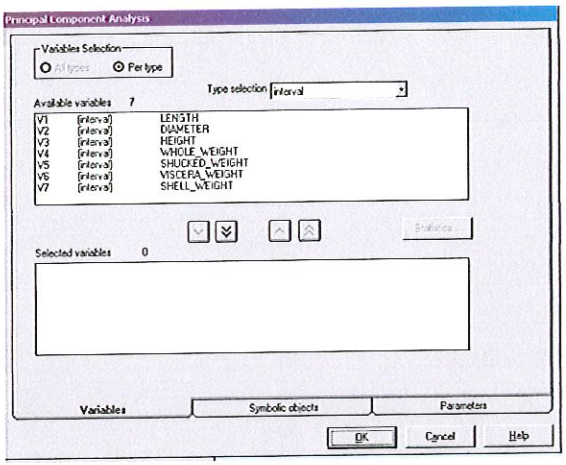


FIG. 5.14 – Choix des variables

Nous pouvons également choisir les individus :

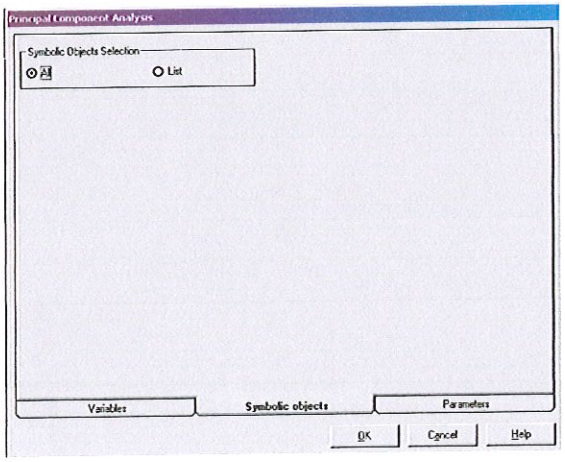


FIG. 5.15 – Choix de la totalité les individus

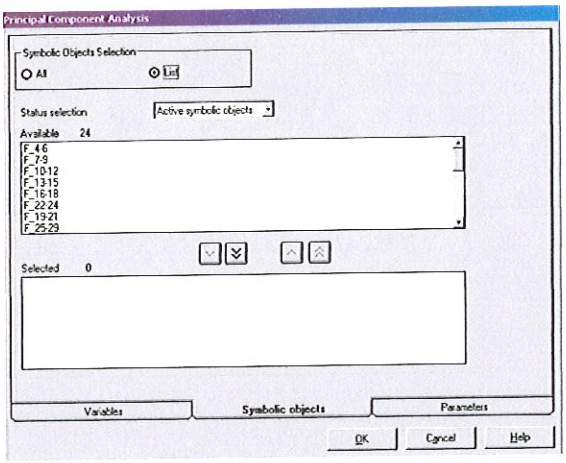


FIG. 5.16 – Choix d'une partie des indivdidus

Le dernier onglet nous permet de choisir les paramètres de la méthode.

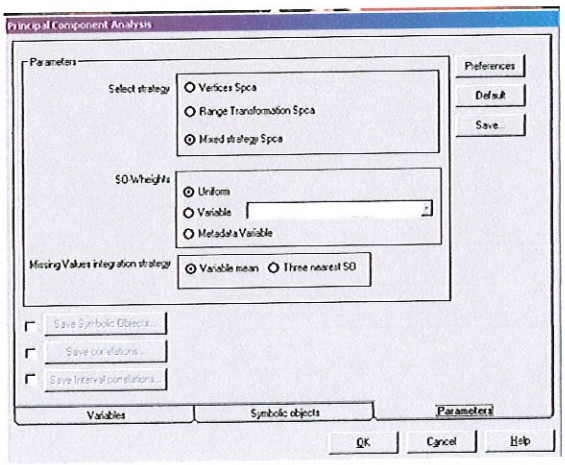


FIG. 5.17 – Choix des paramètres

Nous avons le choix entre trois méthodes. Celle qui nous intéresse ici est la première (`vertices Spca`), il s'agit de la méthode des sommets. La méthode des centres n'est en effet pas reprise dans le logiciel.

Trois méthodes de pondération des individus sont également proposées. Nous avions choisi de donner des poids identiques à tous les individus, ce qui correspond à la première proposition : `SO-weights uniform`.

Nous pouvons également décider de la manière dont d'éventuelles données manquantes seront traitées :

- `variable mean` : elles seront remplacées par un intervalle dont la borne inférieure est la moyenne des bornes inférieures de la variable et la borne supérieure est la moyenne des bornes supérieures
- `three nearest SO` : elles seront remplacées par la moyenne des trois individus les plus proches.

Nous pouvons également sauvegarder dans un fichier sodas :

- les coordonnées des objets symboliques sur le plan formé par les composantes principales (`save symbolic objects`)
- les corrélations des variables symboliques (`save correlation`)
- les corrélations des variables intervalles (`save interval correlation`).

Nous obtenons deux fichiers en sortie : un fichier texte et un fichier graphique.

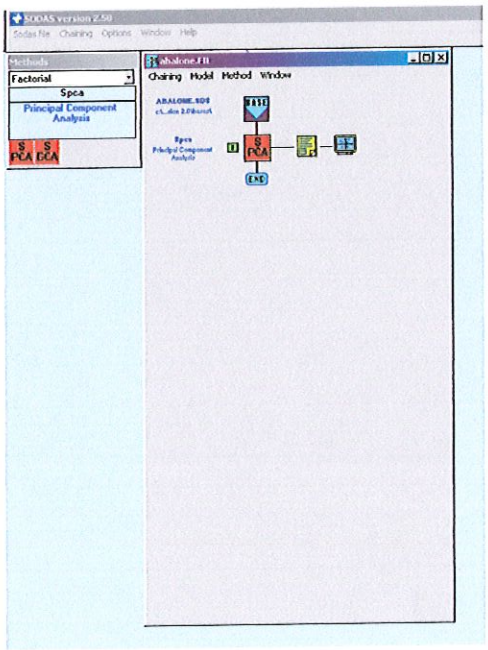


FIG. 5.18 – Exécution de la méthode

La sortie texte

Cette sortie contient :

- la liste des objets et des variables que nous avions sélectionnés
- les valeurs propres de la matrice de covariance, le pourcentage de variance expliqué par chaque valeur propre ainsi qu'un histogramme représentant le pourcentage cumulé de variance expliquée
- les coordonnées des objets sur le plan formé par les composantes principales sous forme d'intervalle
- la contribution absolue des objets aux axes (ce qui correspond à notre indice CTR)
- une mesure de qualité de représentation des objets sur chaque composante principale (c'est-à-dire notre indice COR)
- les corrélations entre les variables originales et les composantes principales
- les corrélations intervalles entre les variables et les composantes principales.

La sortie graphique

Nous obtenons également trois graphiques :

- la représentation des objets par des rectangles (`SOs interval coordinates`)
- le cercle de corrélation : les variables originales sont représentées en fonction de leur corrélation avec les composantes principales sur le cercle unitaire (`Variables circle of correlation`)
- les corrélations intervalles : ce graphique est identique au précédent mais on y représente les corrélations intervalles (`Variables Int Coordinates`).

Remarque : L'analyse en composantes principales classique

Comme nous l'avons vu, l'analyse en composantes principales classique est un cas particulier de l'analyse en composantes principales symbolique.
Pour appliquer la méthode classique, il nous suffit donc d'appliquer la méthode symbolique à un ensemble de données où les données classiques sont représentées par des intervalles où la borne inférieure est égale à la borne supérieure.

5.4 L'analyse factorielle discriminante sous Sodas

Comme pour l'analyse en composantes principales, après avoir choisi notre base de données, nous devons sélectionner notre méthode. Pour l'analyse factorielle discriminante, nous devons aller dans l'onglet `Methods` et sélectionner `Discrimination & regression`. Nous trouvons alors notre méthode : `SFDA`.

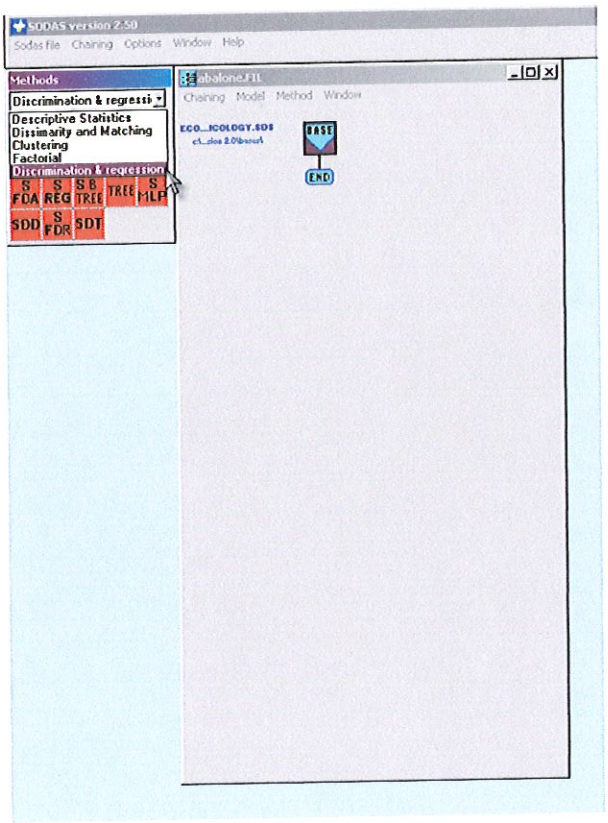


FIG. 5.19 – Choix de la méthode

Lorsque la méthode est sélectionnée, nous pouvons choisir les différents paramètres.

Commençons par choisir nos variables. Ici, il y a une différence. Nous devons en effet définir deux types de variables :

- la variable qui séparera nos individus en différents groupes : `class identifier variable`.
- les variables qui interviendront dans notre analyse : `explanatory`

variables.

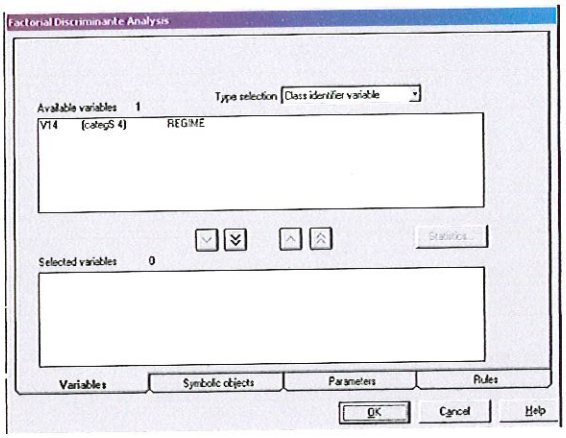


FIG. 5.20 – Choix de la variable classificatoire

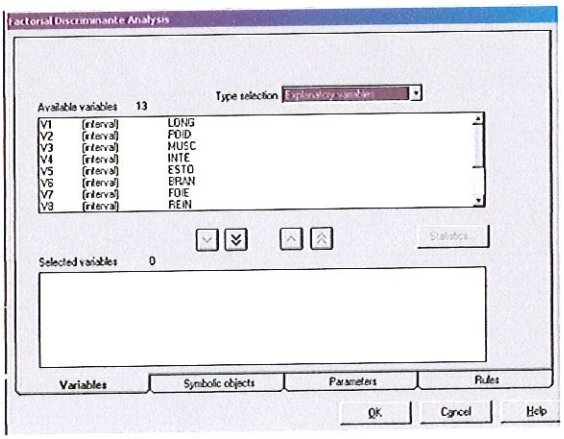


FIG. 5.21 – Choix des variables

Ensuite, nous devons sélectionner les individus qui interviendront dans notre analyse.
De la même façon que pour l'analyse en composantes principales, nous pouvons les sélectionner tous (`All`) ou n'en choisir qu'une partie (`List`).

L'étape suivante consiste à définir les différents paramètres de l'analyse factorielle discriminante symbolique.

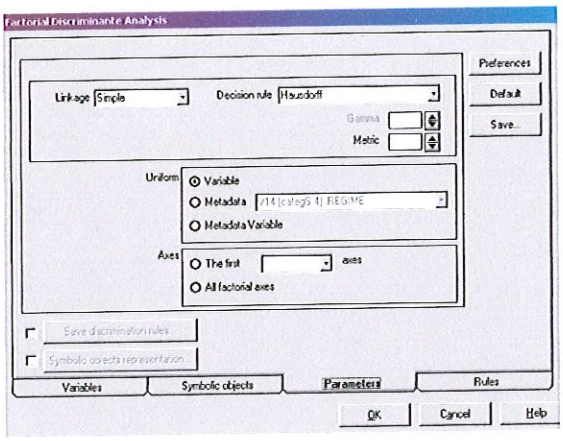


FIG. 5.22 – Choix des paramètres

Nous devons tout d'abord choisir notre règle de décision. Sodas nous en propose trois :

- Ichino-De Carvalho
- l'augmentation du descripteur de potentiel
- Hausdorff.

Dans notre cas, nous utiliserons les deux premières.

Lorsque nous utilisons la distance d'Ichino-De Carvalho (ou celle de Hausdorff), nous pouvons choisir le lien :

- simple : la dissimilarité entre les objets et les classes est calculée en considérant la distance minimum entre le nouvel objet que l'on souhaite classer et tous les objets de l'ensemble E
- moyen : on utilise ici la dissimilarité moyenne entre le nouvel objet et tous ceux de E
- ou complet : la distance considérée est la distance maximale entre le nouvel objet et les objets de E.

Le nouvel objet est affecté à la classe de l'individu dont la dissimilarité est minimale.

Nous avons considéré que les objets avaient des poids identiques, nous utilisons donc la valeur par défaut : `uniform`.

Ici, nous pouvons choisir le nombre d'axes factoriels que nous souhaitons avoir (`the first... axes`) ou les calculer tous (`All factorial axes`).

Nous pouvons également sauvegarder dans un fichier sodas :

- la règle de classification (`save discrimination rule`)
- les coordonnées des objets symboliques sur le plan factoriel (`save symbolic objects representation`).

Lorsque nous avons défini les paramètres, nous pouvons exécuter la méthode. Ici aussi, nous obtenons deux fichiers en sortie : un fichier texte et un fichier graphique.

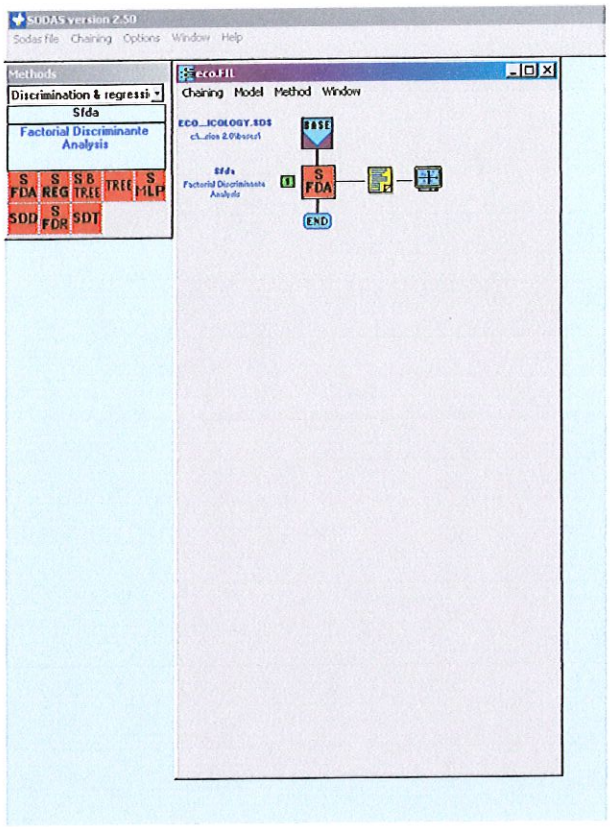


FIG. 5.23 – Exécution de la méthode

La sortie texte

Ce fichier contient :

- l'identifiant des classes et le nom des classes
- les résultats de la procédure de sélection automatique qui assigne un individu à une classe
- la liste des variables que nous avons sélectionnées
- la liste des classes avec les différents objets qui y appartiennent
- les valeurs propres et les vecteurs propres de la matrice $T^{-1}B$
- les coordonnées des individus sur le plan factoriel sous forme d'intervalles
- la matrice de distances entre les différents objets
- la table d'assignation des différents individus aux différentes classes en fonction de la règle de décision que nous avions choisie
- la matrice de classification contenant le nombre d'individus affectés aux classes a priori et le nombre d'individus affectés aux différentes

classes en fonction de notre règle d'affectation
- le taux de bons classements.

La sortie graphique

Nous obtenons la représentation des objets et des classes par des rectangles.

Remarque : L'analyse factorielle discriminante classique

Pour effectuer une analyse factorielle discriminante classique, il nous suffit de prendre une base de données où les données classiques sont représentées par des intervalles où la borne inférieure est égale à la borne supérieure.

Chapitre 6

Application de l'analyse en composantes principales

6.1 L'analyse en composantes principales classique : l'évolution des crânes égyptiens

6.1.1 Présentation des données

Les données consistent en une série de mesures effectuées sur des crânes d'hommes Egyptiens de cinq périodes différentes. Nous avons 150 individus sur lesquels cinq mesures ont été effectuées :

- la largeur maximale que nous noterons MB
- la taille BH
- la longueur BL
- la taille du nez NH
- l'année de formation du crâne $YEAR$

Ces données sont disponibles sur *http ://lib.stat.cmu.edu/DASL/Stories/EgyptianSkullDevelopment.html.*

6.1.2 Choix des paramètres

Pour travailler avec ces données classiques, il faut tout d'abord les transformer en variables intervalles où la borne inférieure est égale à la borne supérieure.

Une fois cela fait, nous pouvons choisir les variables et les individus qui interviendront dans notre analyse ainsi que la méthode utilisée. Nous prendrons ici nos cinq variables, nos 150 individus et nous choisirons la méthode des sommets.

6.1.3 Résultats

Les valeurs propres de la matrice de variance-covariance ainsi que le pourcentage de variance expliquée par chacune d'entre elles sont donnés par :

Eigenvalues	Explained Inertia		%	Cumulated %	Histogram 0-------25%--------50%-------75%-------100%
Ev.1	1.77416	35.48319	35.48319	|**********	
Ev.2	1.25057	25.01138	60.49458	|*******	
Ev.3	0.76114	15.22285	75.71742	|****	
Ev.4	0.71405	14.28110	89.99852	|****	

FIG. 6.1 – Pourcentage de variance expliquée par chaque valeur propre

Aucune des valeurs propres n'explique un pourcentage important de la variance, le maximum étant de 35% pour la première. Nous pouvons voir que si nous ramenons nos individus dans l'espace formé par les deux premières composantes principales, nous n'expliquons que 60, 5% de la variance totale ce qui est peu.

Examinons tout de même les résultats que l'on obtient dans ce plan. Nous obtenons la représentation suivante des 150 individus :

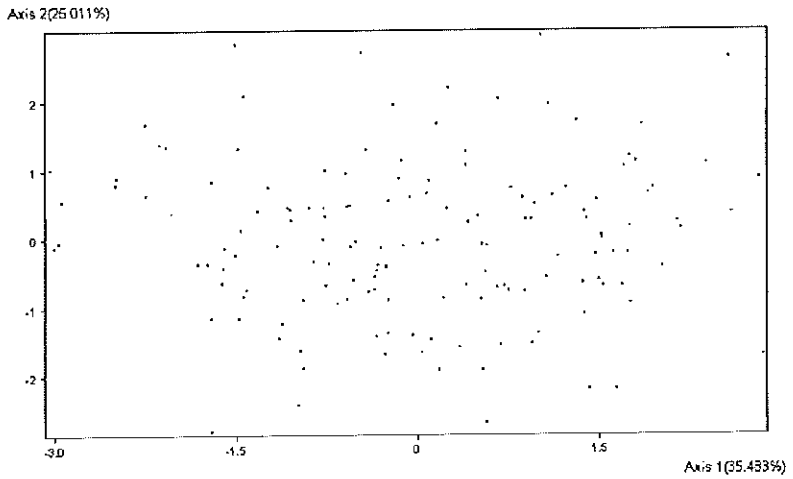


FIG. 6.2 – Représentation des individus

Pour y voir plus clair, représentons les individus en fonction de la période à laquelle ils appartiennent.

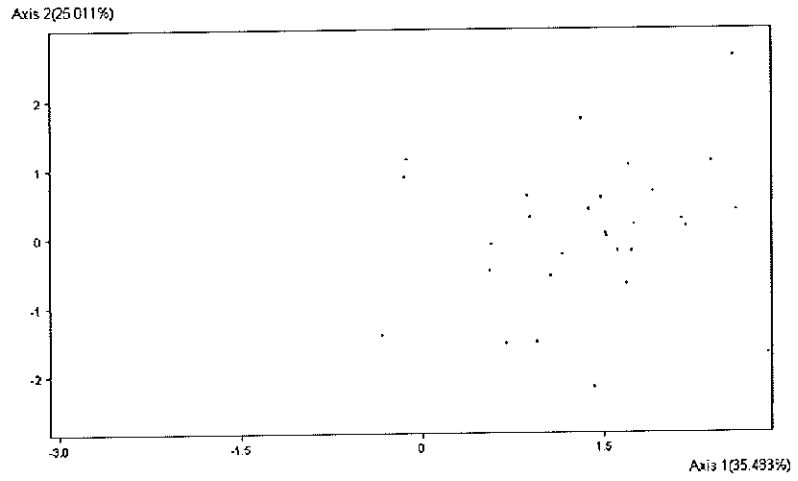


FIG. 6.3 – Représentation des individus dont l'année de formation est -4000

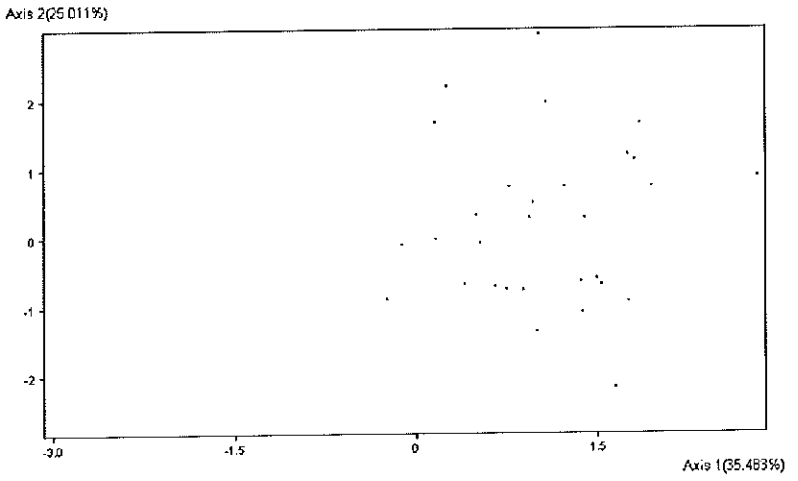


FIG. 6.4 – Représentation des individus dont l'année de formation est -3300

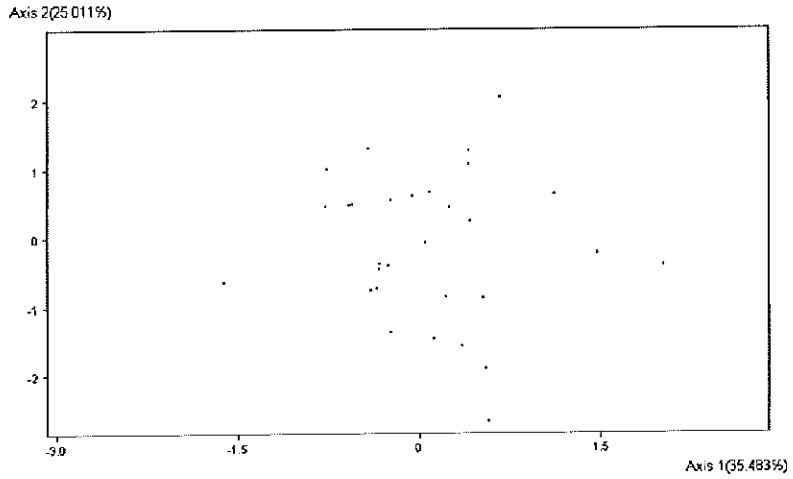


FIG. 6.5 – Représentation des individus dont l'année de formation est -1850

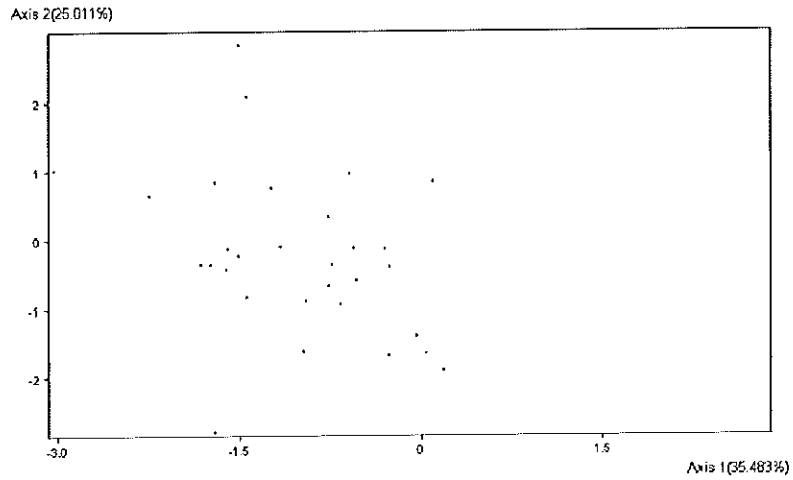


FIG. 6.6 – Représentation des individus dont l'année de formation est -200

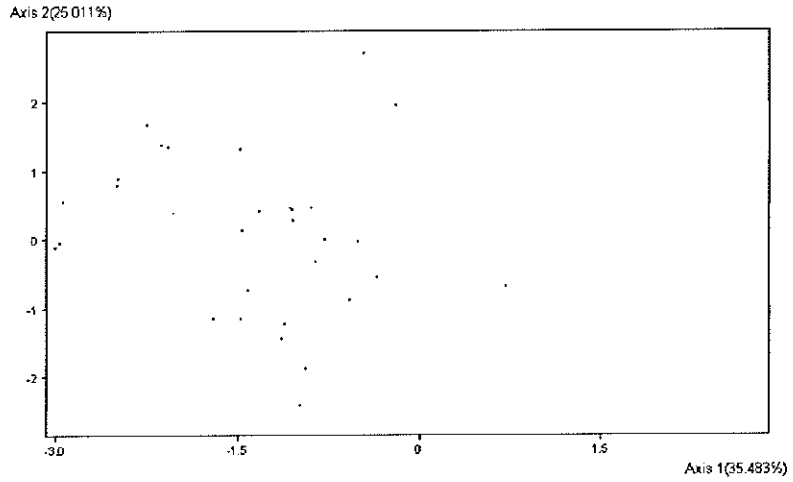


FIG. 6.7 – Représentation des individus dont l'année de formation est 150

Nous pouvons remarquer que plus les périodes passent, plus les individus se déplacent vers la gauche.

6.1.4 Interprétation des résultats

Examinons les corrélations entre les variables et les composantes principales :

Correlations between variables and factors (5 vars,4 fact)=

Var.	Factor 1	Factor 2	Factor 3	Factor 4
MB	-0.61632	-0.37145	0.64419	-0.08188
BH	0.38747	-0.67633	0.00840	0.62640
BL	0.70627	-0.31032	0.34968	-0.37428
NH	-0.25421	-0.74364	-0.47198	-0.38855
YEAR	-0.82507	-0.07666	-0.03251	0.15465

FIG. 6.8 – Corrélations entre les variables et les composantes principales

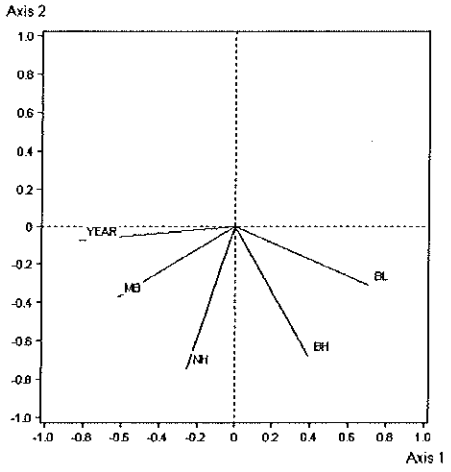


FIG. 6.9 – Cercle de corrélation

La première composante principale est corrélée principalement avec les variables MB, BL et $YEAR$ et la deuxième avec BH et NH. La première composante principale tient donc compte principalement de la forme du crâne, la deuxième explique plutôt le relief du crâne.

Si nous examinons les graphiques par période, nous comprenons donc pourquoi plus les périodes sont lointaines, plus les individus ont une première composante principale importante. De même, plus la largeur du crâne est importante, plus les individus ont une valeur égative pour cette première composante principale. Nous aurons également que les individus dont la longueur du visage est importante auront une première composante principale plus grande.

Les individus ayant une valeur négative pour la première composante principale seront donc des individus provenant d'une période plus récente, dont la largeur du visage est importante et la longueur moins importante. Ceux qui auront une première composante principale positive seront les individus plus anciens dont la largeur du visage est plus faible mais la longueur plus grande.
En examinant les données, nous voyons que la seule variable ($YEAR$ non comprise) qui semble augmenter légérement au cours du temps est la variable MB, la largeur maximale, ce qui explique bien que les individus ont, au fil des années, une valeur négative pour la première composante principale.

Les individus semblent répartis autour de la deuxième composante principale et ce, pour toutes les périodes, ce qui signifie que tous les individus ont des tailles de crâne et de nez équivalente quelle que soit la période mais que ces valeurs sont très variables d'un individu à un autre.

6.2 L'analyse en composantes principales intervalle : la reconnaissance des visages

6.2.1 Présentation des données

La reconnaissance automatique de visage comporte trois phases :

- la phase de description des visages qui consiste à extraire les caractéristiques des visages. Pour ce faire, nous pouvons utiliser une technique géométrique qui consiste en une description des visages par un ensemble de paramètres mesurant une série de distances entre différents points de référence du visage ;

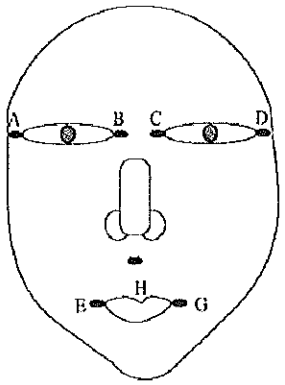


FIG. 6.10 – Points de référence du visage

- la phase de classification qui revient à étudier les visages que l'on a décrits lors de la phase précédente. Elle permet de déterminer les principaux groupes de visages ainsi que les caractéristiques générales de chacun de ces groupes. C'est lors de cette phase que nous utiliserons l'analyse en composantes principales : celle-ci nous permet d'obtenir les images des visages dans un espace de dimension réduite ;

- la phase d'identification qui consiste à comparer les paramètres que l'on obtient avec un nouveau visage et de le comparer avec ceux contenus dans la base des visages qui ont été décrits dans les phases précédentes au moyen de distances.

Notre base de données des visages est constituée d'un ensemble de mesures effectuées sur neuf hommes. Pour chaque personne, nous avons trois séquences d'images.

De par la mobilité du visage, les points de référence et les différentes distances que nous pouvons mesurer varient. Nous décrirons donc ces données par des données intervalles.

La base de données contient 27 séquences d'images (3 par personnes) décrites par 6 variables de type intervalle. Ces variables correspondent aux distances entre deux points de référence du visage : AD, BC, AH, DH, EH et GH.

Ces données proviennent de la thèse de doctorat de A. Chouakria, Extension des méthodes d'analyse factorielle à des données de type intervalle, Université Paris IX-Dauphine, 1998.

6.2.2 Résultats

Nous obtenons les valeurs propres et les pourcentages de variance expliquée suivants :

Eigenvalues	Explained Inertia		%	Cumulated %	Histogram 0-------25%--------50%--------75%--------100%
Ev.1	2.44608	49.24880	49.24880	|**********	
Ev.2	1.56895	31.58889	80.83769	|*******	
Ev.3	0.44357	8.93067	89.76836	|**	
Ev.4	0.25293	5.09238	94.86074	|*	
Ev.5	0.20356	4.09840	98.95914	|	

FIG. 6.11 – Valeurs propres et pourcentages de variance expliquée

Nous prendrons deux composantes principales qui expliqueront 80% de la variance totale des données.

Nous obtenons alors la représentation des 27 séquences d'images suivante :

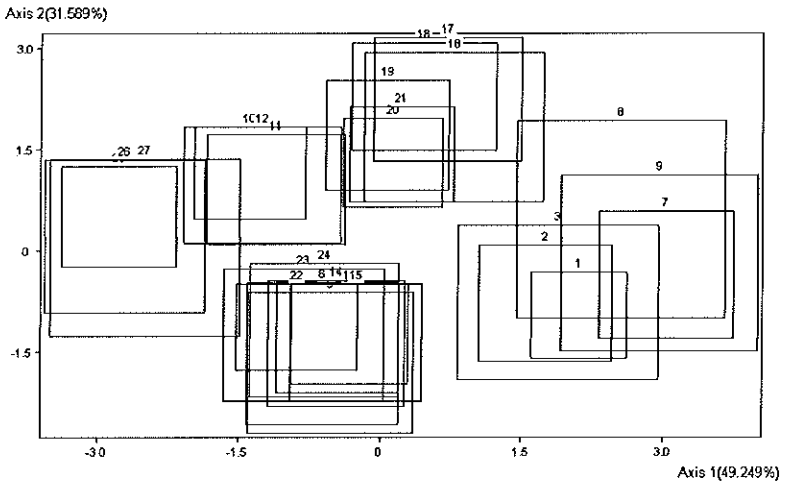


FIG. 6.12 – Représentation des séquences d'images dans le plan formé par les deux premières composantes principales

Si nous isolons les 3 séquences d'images relatives à la même personne nous pouvons voir que ces séquences se regroupent :

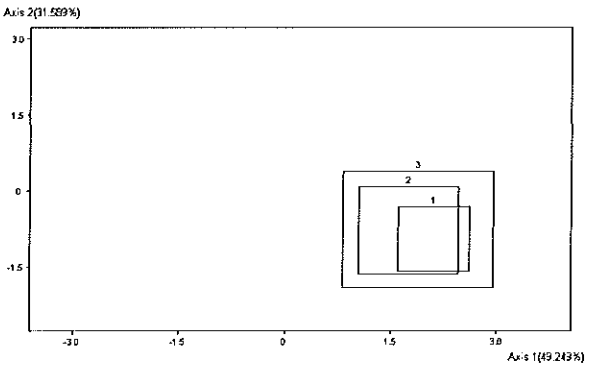


FIG. 6.13 – Séquences d'images relatives à la première personne

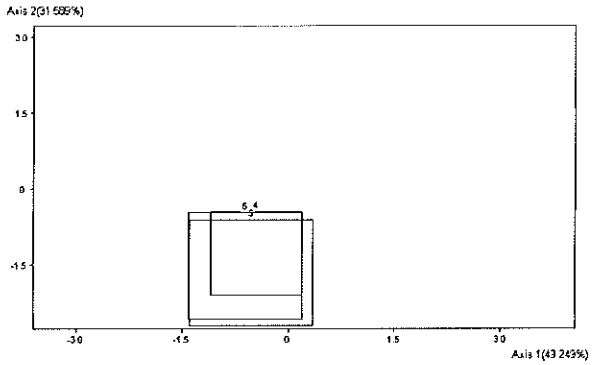


FIG. 6.14 – Séquences d'images relative à la deuxième personne

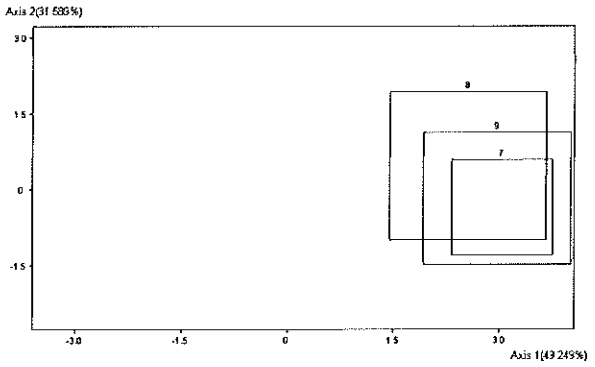


FIG. 6.15 – Séquences d'images relative à la troisième personne

6.2.3 Interprétation des résultats

Examinons les corrélations entre les variables et les composantes principales.

Nous pouvons alors voir que les corrélations les plus fortes apparaissent au niveau de la première composante principale. Cette composante explique en effet 80% de la variance totale des données.

Cette première composante est principalement corrélée avec les variables AH et DH c'est-à-dire avec les distances entre les yeux et le dessus des lèvres. Les corrélations étant négatives, nous aurons que les personnes chez lesquelles cette distance est grande auront une première composante principale négative.

Correlations between variables and factors (6 vars,5 fact)=

Var.	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
AD	-0.41931	-0.26095	0.16823	0.29725	0.47442
BC	-0.18904	-0.17660	0.46709	0.24955	0.32386
AH	-0.65932	0.31450	0.40685	0.25672	0.17077
DH	-0.70037	0.07720	0.04038	0.28524	0.06926
EH	0.61110	-0.11686	0.68688	0.54791	0.43843
GH	0.55305	-0.28279	0.25996	0.49983	0.31220

FIG. 6.16 – Corrélations entre les variables et les composantes principales

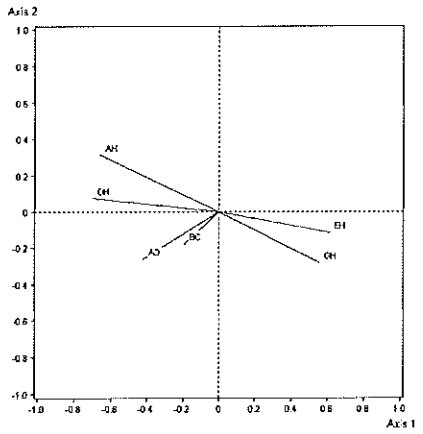


FIG. 6.17 – Cercle des corrélations

La deuxième composante principale n'est fortement corrélée à aucune variable. Les corrélations les moins faibles sont avec les variables AH, GH et AD c'est-à-dire avec la distance entre les yeux, entre l'oeil droit et le dessus des lèvres, et entre le bord des lèvres et le nez. Ces corrélations étant vraiment très faibles nous ne pouvons pas vraiment faire de déductions sur la position des individus par rapport à cette composante principale.

Chapitre 7

Applications de l'analyse factorielle discriminante

7.1 L'analyse factorielle discriminante classique : pédagogie des mathématiques

7.1.1 Présentation des données

Dans le cadre d'une étude sur la pédagogie des mathématiques, 20 personnes ont passé différentes épreuves :

- une épreuve combinatoire ($COMB$)
- une épreuve de probabilité ($PROB$)
- une épreuve de logique ($LOGI$)
- une épreuve de mathématique ($MATH$)
- et une épreuve collective ($EPCO$).

Nous avons également mesuré le QI de ces individus.

Les personnes sont réparties en 3 classes en fonction des résultats qu'ils ont obtenus à leur épreuve en mathématique. Ils appartiennent à :

- C_1 s'ils ont eu moins de 10
- C_2 s'ils ont eu entre 10 et 13
- C_3 s'ils ont eu 14 ou plus.

Les classes C_1, C_2 et C_3 contiennent respectivement $7, 8$ et 5 individus.

Ces données sont disponibles sur :
http ://piaget.psycho.univ-paris5.fr/statistiques/donnees/psychom.htm.

7.1.2 Résultats

Les valeurs propres de la matrice $T^{-1}B$ sont données par :

Eigenvalues	Inertia	Percentage of expl. inertia	Cumulated % of inertia	Histogram 0----------50%----------100%
1	0.09104	70.695	70.695	|******************************
2	0.02826	21.941	92.636	|********
3	0.00948	7.364	100.000	|**

FIG. 7.1 – Valeurs propres et pourcentage de variance expliquée

Nous obtenons donc 2 axes factoriels qui expliquent 92% de la variance totale des données.

7.1.3 Règle d'affectation

En utilisant la distance d'Ichino-Carvalho pour affecter nos individus, nous obtenons le taux de bons classements suivant :

Classification matrix
the rows are the apriori classes
the column are the assigned classes

From \ To|	Class 1	Class 2	Class 3
Class 1|	2.333	2.333	2.333
Class 2|	2.667	2.667	2.667
Class 3|	1.667	1.667	1.667

Correct classification ratio ===>33.333 %

FIG. 7.2 – Taux de bons classements par la règle d'Ichino-Carvalho

Ce taux de 33% est très faible. Essayons la règle d'affectation basée sur l'augmentation minimale du descripteur de potentiel.

Classification matrix
the rows are the apriori classes
the column are the assigned classes

From \ To|	Class 1	Class 2	Class 3
Class 1|	5.000	1.000	1.000
Class 2|	-	7.000	1.000
Class 3|	1.000	1.000	3.000

Correct classification ratio ===>75.000 %

FIG. 7.3 – Taux de bons classements par le *pdi*

Nous obtenons alors un taux de bons classements de 75% ce qui est nettement mieux.

Nous obtenons la représentation des individus sur le plan factoriel :

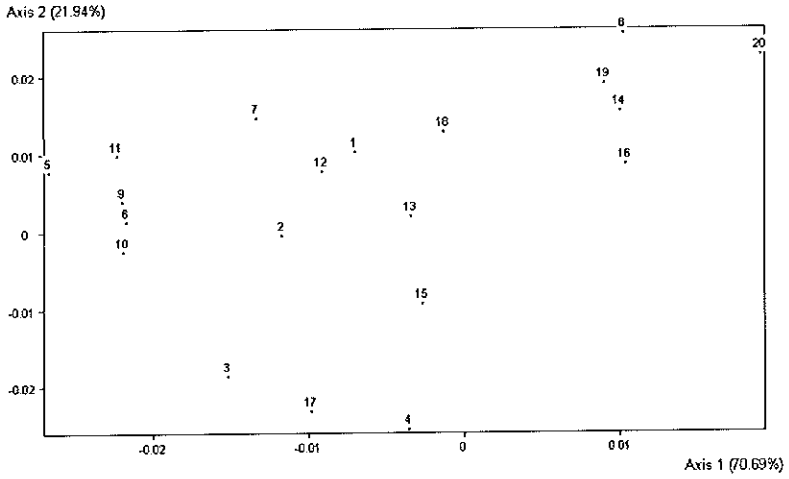


FIG. 7.4 – Les individus sur le plan factoriel

ainsi que la représentation des classes :

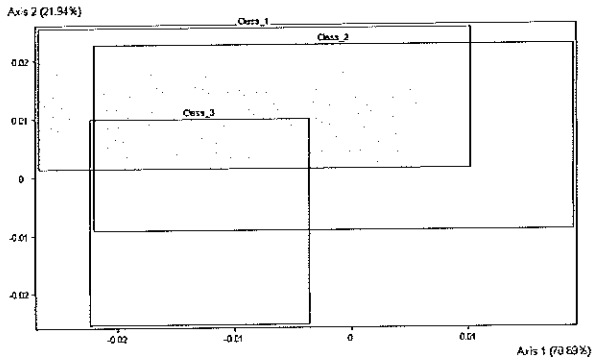


FIG. 7.5 – Première classe

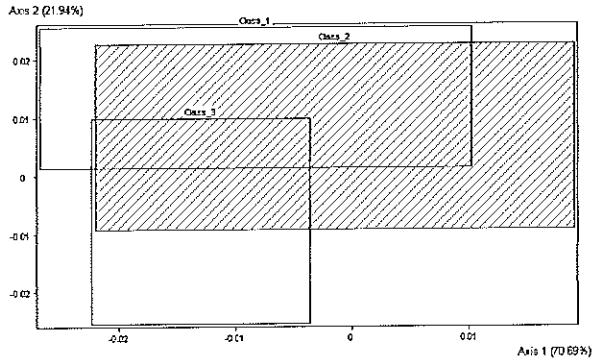


FIG. 7.6 – Deuxième classe

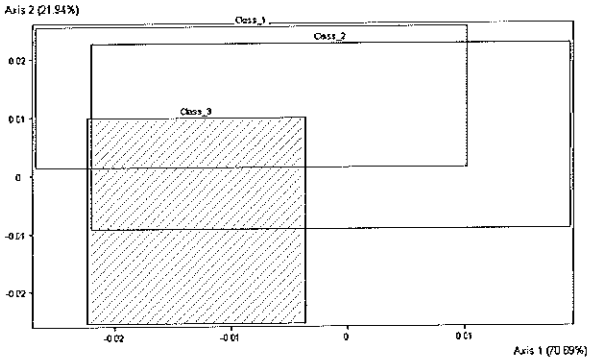


FIG. 7.7 – Troisième classe

7.1.4 Commentaires

Si nous choisissons la distance d'Ichino-Carvalho, nous obtenons de très mauvais résultats. En effet, si nous regardons la répartition des points dans leurs classes d'appartenance a priori, nous pouvons voir que ces classes ne sont pas bien séparées.

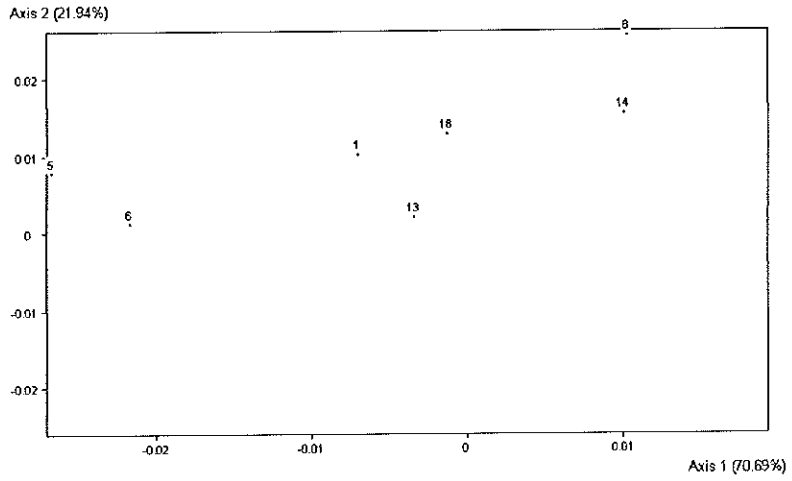


FIG. 7.8 – Individus de la première classe

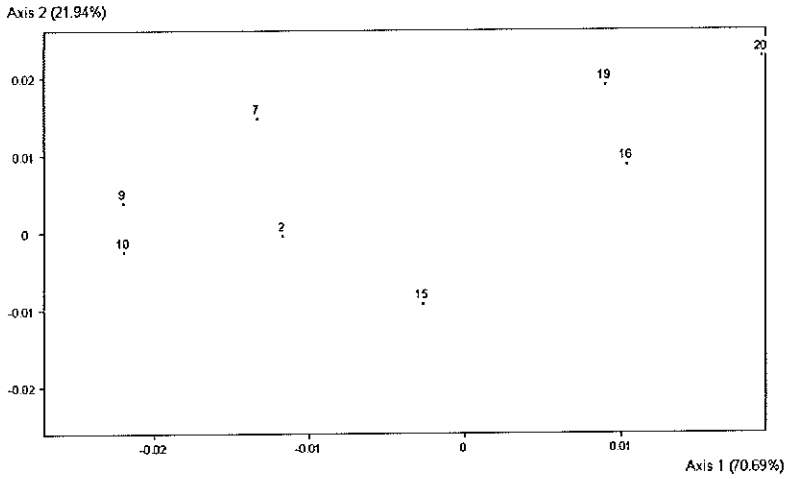


FIG. 7.9 – Individus de la deuxième classe

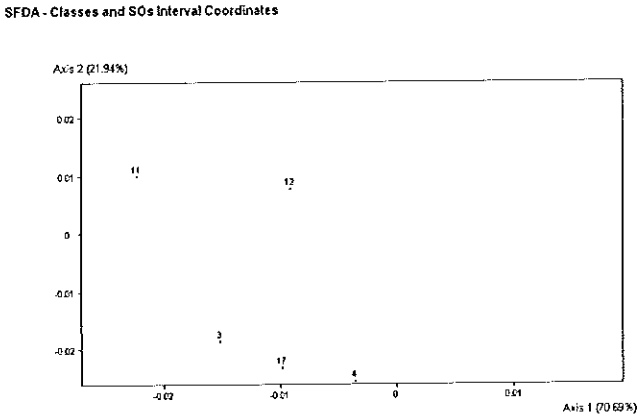


FIG. 7.10 – Individus de la troisième classe

Les distances entre les différents objets sont toutes égales quelque soit la valeur que nous choisissons pour γ et quelque soit le lien choisi. De par ce fait, les objets se retrouvent classés aléatoirement comme le montre la matrice de classification que l'on obtient. Les objets sont affectés à une des trois classes avec une probabilité de 0.3.

Assignment Matrix the rows are the objects the columns the classes
in brakets () the apriori class of the misclassified object

	Class 1	Class 2	Class 3
1	0.3	0.3(1)	0.3(1)
2	0.3(2)	0.3	0.3(2)
3	0.3(3)	0.3(3)	0.3
4	0.3(3)	0.3(3)	0.3
5	0.3	0.3(1)	0.3(1)
6	0.3	0.3(1)	0.3(1)
7	0.3(2)	0.3	0.3(2)
8	0.3	0.3(1)	0.3(1)
9	0.3(2)	0.3	0.3(2)
10	0.3(2)	0.3	0.3(2)
11	0.3(3)	0.3(3)	0.3
12	0.3(3)	0.3(3)	0.3
13	0.3	0.3(1)	0.3(1)
14	0.3	0.3(1)	0.3(1)
15	0.3(2)	0.3	0.3(2)
16	0.3(2)	0.3	0.3(2)
17	0.3(3)	0.3(3)	0.3
18	0.3	0.3(1)	0.3(1)
19	0.3(2)	0.3	0.3(2)
20	0.3(2)	0.3	0.3(2)

FIG. 7.11 – Matrice de classification obtenue par la distance d'Ichino-Carvalho

7.1.5 Représentation des différentes classes

Représentons les différentes individus suivant leurs classes à l'aide des zoom stars afin d'obtenir certaines informations concernant les caractéristiques principales de ces classes.

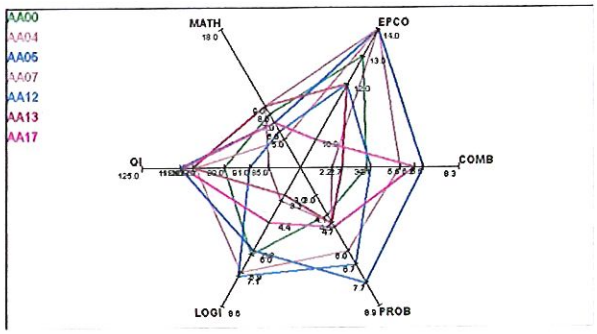


FIG. 7.12 – Représentation des individus de la première classe

La première classe correspond aux individus ayant obtenu moins de 10 à leur épreuve en mathématique.

Nous pouvons voir que les résultats obtenus aux autres tests sont très variables mais que leurs QI se situe généralement entre 100 et 110.

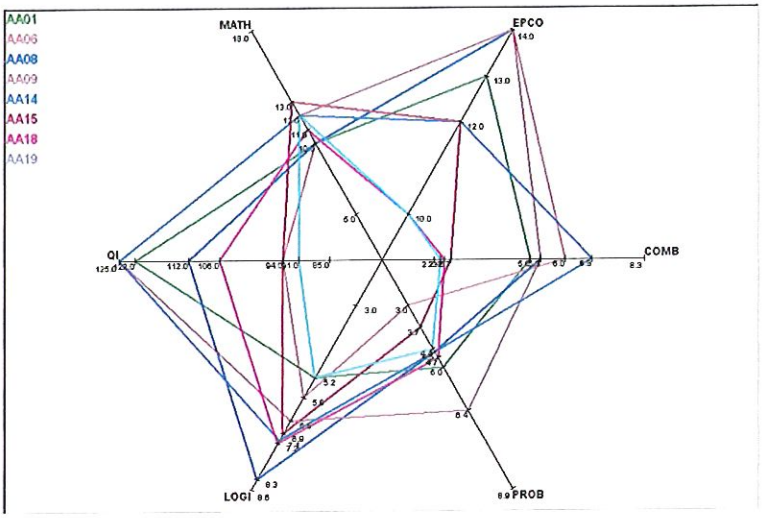


FIG. 7.13 – Représentation des individus de la deuxième classe

Cette classe représente les individus ayant obtenu entre 10 et 13 à leur épreuve en mathématique. Nous pouvons observer plusieurs choses :

- la note que les individus ont obtenue à leur épreuve en calcul combinatoire se regroupe en deux zones : autour de 2 et entre 5 et 7.
- la majorité des résultats obtenus à l'épreuve de probabilité se trouvent entre 4 et 5
- leurs résultats en logique sont entre 5 et 7.

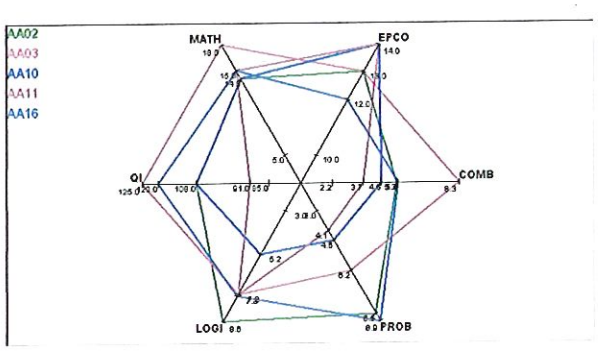


FIG. 7.14 – Représentation des individus de la troisième classe

La dernière classe représente les individus ayant obtenus plus de 14 à leur épreuve en mathématique. Ces individus ont généralement de mauvais résultats à l'épreuve combinatoire (entre 3 et 5) mais de très bon résulats en logique (plus de 7).

De toutes ces observations, nous pouvons déduire que :

- les résultats en logique sont meilleurs dans la classe 3, plus les individus réussissent leur épreuve de mathématiques, plus ils sont logiques.
- les individus ayant un QI plus élevé réussissent leur épreuve en mathématiques (ils obtiennent plus de 10 c'est-à-dire qu'ils appartiennent à la classe 2 ou 3).

7.2 L'analyse factorielle discriminante symbolique : musique

7.2.1 Présentation des données

Le jeu de données est constitué de 33 chanteurs. Pour chacun de ces chanteurs nous avons les renseignements suivants :

- sa nationalité, le pays et la région dans lesquels il habite
- sa date de naissance
- le genre de sa musique
- le nombre de récompenses qu'il a reçues
- le nombre d'albums qu'il a enregistré
- plusieurs titres de chansons qui ont reçu une récompense
- la durée de ses chansons
- le nombre de ventes de ses chansons ainsi que les récompenses que cette chanson a obtenues
- la date de sortie de ses chansons ainsi que le nom et la date de sortie de l'album sur lequel elles se trouvent
- l'éditeur et le distributeur de cet album.

Ces renseignements sont représentés par 16 variables symboliques : 11 variables modales, 5 variables intervalles, et 1 variable multivaluée.

Les chanteurs sont répartis a priori en différents groupes suivant leur nationalité. Nous aurons les 9 classes suivantes :

- C_1 les chanteurs Américains
- C_2 les chanteurs Français
- C_3 les chanteurs Australiens
- C_4 les chanteuses Belges
- C_5 les chanteuses Françaises
- C_6 les chanteuses Canadiennes
- C_7 les chanteuses Américaines
- C_8 les chanteuses Colombiennes
- C_9 les chanteurs Anglais

Ces données sont disponibles sur *http ://www.ceremade.dauphine.fr/touati/sodas-pagegarde.htm.*

7.2.2 Résultats

Les valeurs propres de notre analyse factorielle sont données par :

Eigenvalues	Inertia	Percentage of expl. inertia	Cumulated % of inertia	Histogram 0----------50%----------100%
1	17.54219	99.770	99.770	|**********
2	0.04040	0.230	100.000	|

FIG. 7.15 – Valeurs propres et pourcentages de variance expliquée

Nous aurons donc 2 axes factoriels. Mais nous pouvons voir qu'un seul axe factoriel suffirait puisqu'il représente presque la totalité de la variance des données.

Les individus sont représentés de la manière suivante :

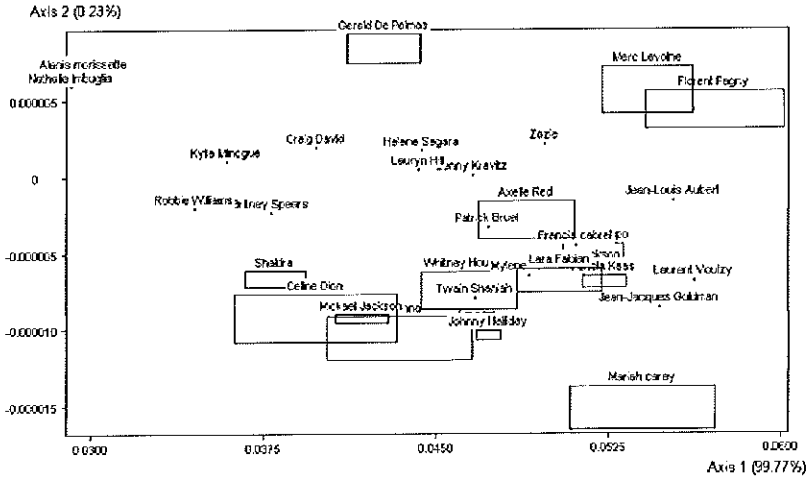


FIG. 7.16 – Représentation des individus sur le plan factoriel

7.2.3 Règle d'affectation

En utilisant la règle d'affectation basée sur l'augmentation minimale du descripteur de potentiel, nous obtenons le taux de bons classements suivant :

Classification matrix
the rows are the apriori classes
the column are the assigned classes

From \ To	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9
Class 1	-	-	-	-	-	-	-	-	5.000
Class 2	-	-	-	-	-	-	-	-	10.000
Class 3	-	-	0.500	-	-	-	-	-	1.500
Class 4	-	-	-	0.500	-	-	-	-	0.500
Class 5	-	-	-	-	-	-	-	-	4.000
Class 6	-	-	-	-	-	-	-	-	4.000
Class 7	-	-	-	-	-	-	-	-	5.000
Class 8	-	-	-	-	-	-	-	0.500	0.500
Class 9	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111

Correct classification ratio ===>4.882 %

FIG. 7.17 – Taux de bons classements

Ce taux de bons classements est très faible. Nous pouvons voir d'après la matrice de classification que presque tous les individus sont affectés à la neuvième classe. La raison est simple : la classe 9 n'est composée que d'un seul individu. La classe est donc représentée par un point et de par ce fait, l'augmentation du descripteur de potentiel sera généralement plus faible avec le seul individu appartenant à C_9.

Si nous retirons cet individu (et donc la classe C_9), nous pouvons voir que nous obtenons de meilleurs résultats (un taux de bons classements d'environ 39%) même si ces résultats ne sont toujours pas très satisfaisants.

Enlever un individu nous a fait gagner 35% de bons classements.

Assignment Matrix the rows are the objects the columns the classes
in brakets () the apriori class of the misclassified object

	Class	1 Class	2 Class	3 Class	4 Class	5 Class	6 Class	7 Class	8 Class	9
Eminem	-	-	-	-	-	-	-	-		1.0(1)
Madonna	-	-	-	-	-	-	-	-		1.0(1)
Mickael Ja	-	-	-	-	-	-	-	-		1.0(1)
Janet Jack	-	-	-	-	-	-	-	-		1.0(7)
Britney Sp	-	-	-	-	-	-	-	-		1.0(7)
Mariah car	-	-	-	-	-	-	-	-		1.0(7)
Lenny Krav	-	-	-	-	-	-	-	-		1.0(1)
Whitney Ho	-	-	-	-	-	-	-	-		1.0(7)
Lauryn Hil	-	-	-	-	-	-	-	-		1.0(7)
Craig Davi	-	-	-	-	-	-	-	-		1.0(1)
Florent Pa	-	-	-	-	-	-	-	-		1.0(2)
Jean-Louis	-	-	-	-	-	-	-	-		1.0(2)
Jean-Jacqu	-	-	-	-	-	-	-	-		1.0(2)
Marc Lavoi	-	-	-	-	-	-	-	-		1.0(2)
Laurent Vo	-	-	-	-	-	-	-	-		1.0(2)
Zazie	-	-	-	-	-	-	-	-		1.0(5)
Johnny Hal	-	-	-	-	-	-	-	-		1.0(2)
Pascal Obi	-	-	-	-	-	-	-	-		1.0(2)
Helene Seg	-	-	-	-	-	-	-	-		1.0(5)
Francis ca	-	-	-	-	-	-	-	-		1.0(2)
Paricia Ka	-	-	-	-	-	-	-	-		1.0(5)
Nathalie I	-	-	0.5	-	-	-	-	-		0.5(3)
Kylie Mino	-	-	-	-	-	-	-	-		1.0(3)
Shakira	-	-	-	-	-	-	-	0.5		0.5(8)
Celine Dio	-	-	-	-	-	-	-	-		1.0(6)
Mylene Far	-	-	-	-	-	-	-	-		1.0(5)
Twain Shan	-	-	-	-	-	-	-	-		1.0(6)
Alanis mor	-	-	-	-	-	-	-	-		1.0(6)
Patrick Br	-	-	-	-	-	-	-	-		1.0(2)
Robbie Wil	0.1(9)	0.1(9)	0.1(9)	0.1(9)	0.1(9)	0.1(9)	0.1(9)	0.1(9)	0.1	
Axelle Red	-	-	-	0.5	-	-	-	-		0.5(4)
Lara Fabia	-	-	-	-	-	-	-	-		1.0(6)
Gerald De	-	-	-	-	-	-	-	-		1.0(2)

FIG. 7.18 – Matrice de classification pour les 33 chanteurs

From \ To| Class	1 Class	2 Class	3 Class	4 Class	5 Class	6 Class	7 Class	8
Class 1|	-	-	-	-	-	5.000	-	-
Class 2|	-	8.000	-	-	-	2.000	-	-
Class 3|	-	-	2.000	-	-	-	-	-
Class 4|	-	-	-	0.500	-	0.500	-	-
Class 5|	-	2.000	-	-	-	2.000	-	-
Class 6|	-	1.000	1.000	-	-	2.000	-	-
Class 7|	-	3.000	-	-	-	2.000	-	-
Class 8|	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125

Correct classification ratio ===>39.453 %

FIG. 7.19 – Taux de bons classements sur les 32 individus

Et nous obtenons des classes différentes.

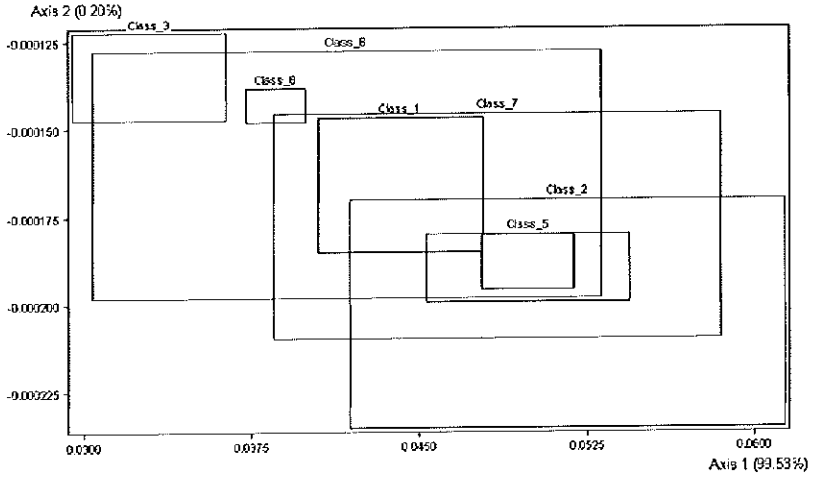


FIG. 7.20 – Représentation des classes obtenues sur le plan factoriel

7.2.4 Commentaires

Le fait que nous obtenions un mauvais taux de bons classements peut toutefois s'expliquer. En effet, lorsque l'on examine les classes que l'on obtient, nous pouvons remarquer que certaines classes sont inclues dans d'autres :

- les classes C_4 et C_5 sont dans la classe C_2

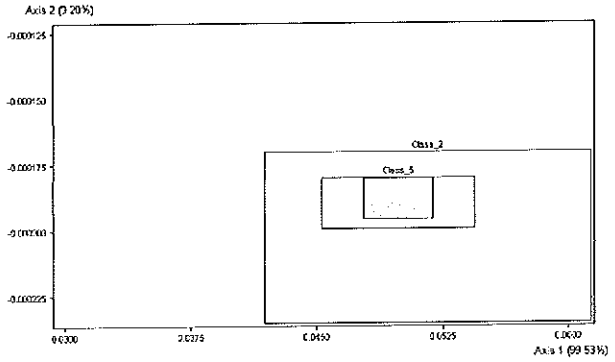


FIG. 7.21 – Représentation des classes C_2, C_4 et C_5

– la classe C_1 est inclue dans la classe C_7

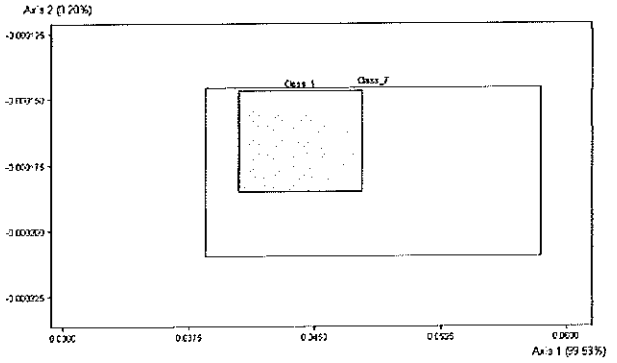


FIG. 7.22 – Représentation des classes C_1 et C_7

– la classe C_8 est inclue dans la classe C_6 et est en partie dans la classe C_7.

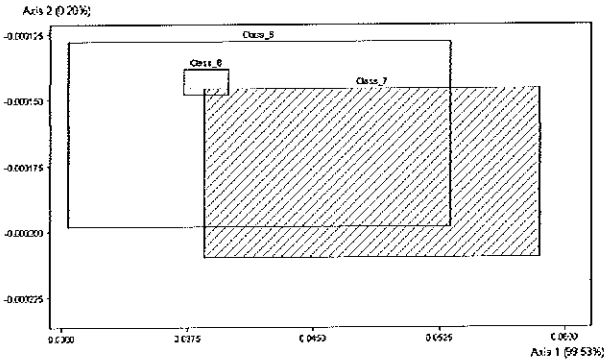


FIG. 7.23 – Représentation des classes C_6, C_7 et C_8

Ces emboitements de classes s'expliquent par deux constatations.
Tout d'abord, aucune des variables que nous utilisons ne nous permet de distinguer les chanteurs des chanteuses ; il est donc logique que les classes C_1 et C_7 (les Américains) ainsi que les classes C_2 et C_5 (les Français) se retrouvent regroupées puisqu'elles correspondent à un même pays.

La deuxième constation vient du fait que a priori, trois classes ne comportent qu'un seul élément : C_9 (que nous avons éliminée), C_4 et C_8. Ces classes diminuent donc le taux de bons classements global. Pourtant, même si on les retire de notre analyse, le taux de bons classements n'augmente pas. En effet, ces classes se retrouvent groupées avec d'autres : C_4 (les Belges) avec les Français et C_8 (les Colombiennes) avec les Américaines (C_7) et les Canadiennes (C_6) puisqu'elles sont toutes du continent américain.

La classe C_9 (les Anglais) ne pouvait pas être regroupée de cette façon, c'est pour cela que son élimination nous a permis d'augmenter le taux de bons classements.

Conclusion

Nous venons de voir deux méthodes qui nous permettent de réduire le nombre de variables intervenant dans la description de différents individus : l'analyse en composantes principales et l'analyse factorielle discriminante.

Grâce à ces méthodes, nous pouvons obtenir une représentation graphique des individus dans un espace de dimension 2 ou 3, quelque soit le nombre de variables de départ.

Ces deux méthodes se différencient par la manière dont nous allons chercher à réduire cet espace. L'analyse en composantes principales cherche à le réduire en tenant compte de la structure globale des données c'est-à-dire de leur variance totale alors que l'analyse factorielle discriminante cherche à le diminuer en tentant de conserver au mieux une éventuelle structure en classes que ces données pourraient avoir. Il nous faut alors considérer les variances inter et intra-classes.

L'analyse en composantes principales symbolique ne concerne que les variables intervalles. Les individus sont donc directement représentés par des hyperrectangles. Ces hyperrectangles peuvent être définis par leurs centres ou par leurs sommets, ce qui donne lieu aux deux méthodes que nous avons abordées :

- la méthode des centres qui correspond à une analyse inter-classes
- la méthode des sommets qui correspond à une analyse inter et intra-classes.

Ces méthodes consistent à appliquer une analyse en composantes principales classique soit aux centres, soit aux sommets des hyperrectangles, et à construire les composantes principales intervalles à l'aide des composantes principales classiques. Les individus sont alors représentés par des hyperrectangles dans le plan formé par les composantes principales.

Pour évaluer la qualité des résultats que nous obtenons, nous avons défini plusieurs paramètres. Ceux-ci nous permettent de voir si les individus sont bien représentés sur le plan formé par les composantes principales. L'interprétation des composantes principales se fait quant à elle par l'étude des corrélations entre les variables originales et ces composantes principales.

L'analyse factorielle discriminante symbolique est applicable à tous les types de données symboliques. Afin de travailler avec toutes ces variables simultanément, nous avons mis en place un système de codage qui nous permet de représenter les individus par des hyperrectangles. Mais nous devons conserver le caractère compact des variables symboliques ce qui donne lieu à une étape supplémentaire : la quantification des sommets des hyperrectangles. Nous appliquons alors l'analyse factorielle discriminante classique à ces sommets quantifiés. Nous obtenons alors une représentation des individus et des classes par des hyperrectangles.

De par le caractère discriminant de cette méthode, nous devons construire une règle d'affectation nous permettant d'affecter les individus aux différentes classes.

Ici, l'évaluation de la qualité des résultats se fera par l'estimation des taux de bons et de mauvais classements par des méthodes comme l'échantillon test, le bootstrap ou encore le leave-one-out.

Nous avons terminé l'étude de ces méthodes par une série d'applications réalisées avec le logiciel Sodas, ce qui nous a permis de mettre en évidence certaines limites de ces méthodes ou d'une étape particulière de la méthode. Nous n'obtenons en effet pas toujours de bons résultats. Il peut arriver que 2 ou 3 composantes principales ou axes factoriels ne soient pas suffisants pour obtenir une bonne représentation des individus. Il peut aussi arriver, avec l'analyse factorielle discriminante, que la règle de classification que nous avons définie ne nous donne pas un bon taux de classement.

Dans ce mémoire, nous avons étudié les méthodes les plus classiques mais d'autres versions de ces méthodes ont été mises au point afin de tenir compte de certaines particularités des données comme l'analyse en composantes principales classique normée, l'analyse en composantes principales sous contraintes, ...
Certaines étapes ont également été revues. On trouve ainsi différents types de représentations ou de codage des données, d'autres techniques de validation, d'autres règle d'affectation, ...

L'étude des méthodes factorielles ne s'arrête donc pas ici, les bases sont juste posées afin de nous permettre d'aller plus loin dans ce domaine.

□

Bibliographie

[1] Cdrom, 2002. Ecole SODAS, PORTO, Factorial analysis of symbolic objects.

[2] Anonyme. www.wu-wien.ac.at/usr/h99c/h995/826/distance.pdf. Mahalanobis distance.

[3] M. Bardos. *Analyse discriminante, application au risque et scoring financier.* Dunod, Paris, 2001.

[4] Ceremade. http ://www.ceremade.dauphine.fr/ touati/sodas-pagegarde.htm.

[5] A. Chouakria. *Extension des méthodes d'analyse factorielle à des données de type intervalle.* Université Paris IX-Dauphine, 1998. Thèse de doctorat.

[6] DASL. http ://lib.stat.cmu.edu/dasl/stories/europeanjobs.html. European jobs data.

[7] Asso developers and partners. User manual for sodas 2 software. http ://www.info.fundp.ac.be/asso, 2004.

[8] F. Dubeau and J. Savoie. http ://www.labmath.uqam.ca/ annales/16-1/125.html. De l'interpolation à l'aide d'une fonction spline définie sur une partition quelconque.

[9] H.-H. Bock et E. Diday. *Analysis of symbolic data, exploratory methods for extracting statistical information from complex data.* Springer, Berlin, 2000.

[10] FUNDP. Help guide for sodas 2 software. http ://www.info.fundp.ac.be/asso, 2004.

[11] D. Garson. http ://www2.chass.ncsu.edu/garson/pa765/factor.htm. Factor analysis.

[12] A. Hardy. *Statistique.* Librairie des sciences, Université de Namur, 2003.

[13] A. Hardy. Aspects statistiques de la classification, 2005.

[14] R. Boné M. Crucianu, J.-P. Asselin de Beauville. *Méthodes factorielles pour l'analyse des données, méthodes linéaires et extensions non-linéaires.* Lavoisier, 2004.

[15] G. Saporta. *Probabilité, analyse des données et statistiques.* Technip, 1990.

[16] The species iris group of north America. http ://www.badbear.com/signa/signa.pl.

[17] Statsoft. http ://www.statsoft.com/textbook/stcoran/html. Correspondance analysis.

[18] Statsoft. http ://www.statsoft.com/textbook/stfacan.html. Principal components and factor analysis.

[19] L. Swysen and G. Seret. *Atlas Erasme, espace et société.* Erasme, 1996.

[20] R. Verde. *Analyse des données symboliques.* Université de Namur, 2006.

Annexes

Annexe I : Création d'une base de données Access

Pour obtenir une base de données symboliques utilisable par Sodas, c'est-à-dire une base au format `.sds`, nous devons tout d'abord construire une base de données access et ensuite utiliser cette base avec le logiciel DB2SO qui est inclu dans Sodas.

Quelques définitions

Une **base de donnée relationnelle** est un ensemble de données, représentées sous forme de tables, entre lesquelles il existe des relations.
Dans le cas qui nous intéresse, les **colonnes** de ces tables représenteront les variables et les **lignes** les individus.

Dans ces tables, nous devrons définir une variable comme **clé primaire**. Il s'agit d'un identifiant qui nous permet de représenter les différentes lignes de la table. Cette variable aura donc des valeurs distinctes.
Dans notre cas, il s'agira de l'identifiant des individus.

Création d'une nouvelle base de données

Pour créer une nouvelle base de données access, il suffit de selectionner `Nouvelle base de données` dans l'onglet `fichier`.

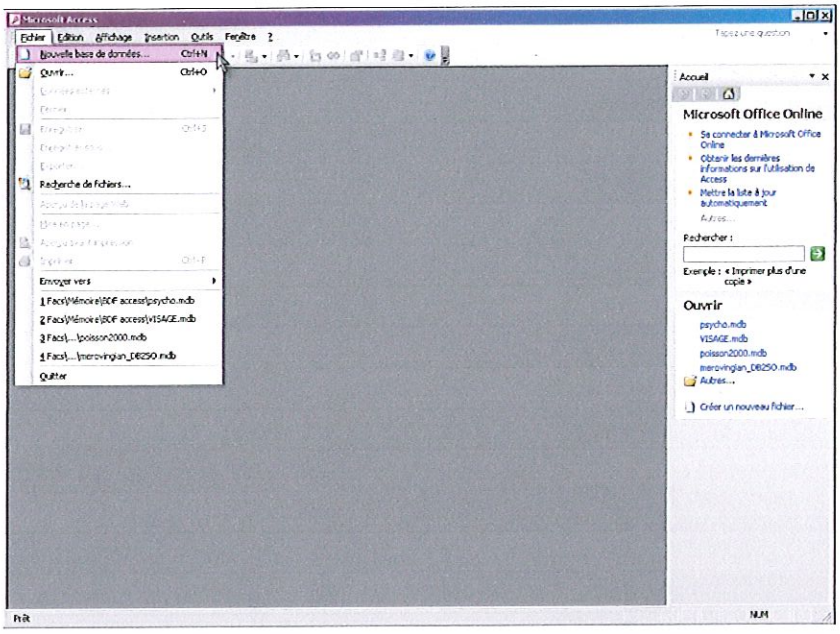


FIG. 7.24 – Création d'une nouvelle base de données

Nous pouvons alors choisir de partir d'une base de données vide ou d'une base de données existante. Partons d'une base de données vide.

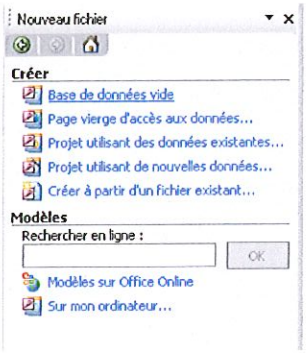


FIG. 7.25 – Options pour la création d'une base de données

Nous devons ensuite nommer (avec l'extension `.mdb`) et donner l'emplacement de notre base de données (ici, nous aurons `exemple.mdb`). Elle est alors créée et nous pouvons construire les différentes tables la constituant ainsi que les requêtes qui nous permettront de définir les liaisons entre ces tables.

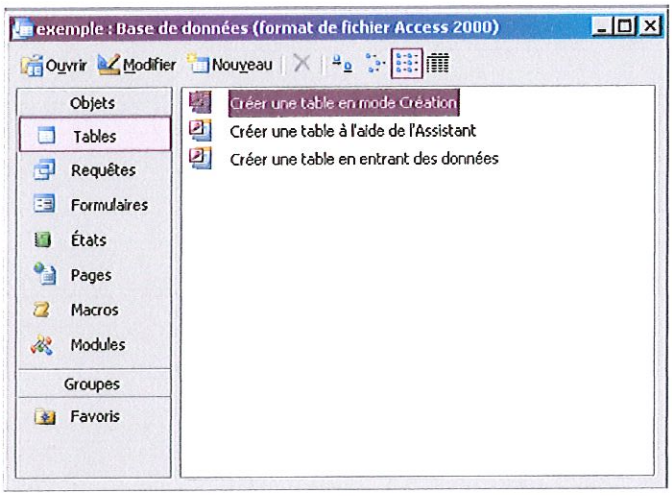


FIG. 7.26 – La base de données `exemple.mdb`

Création de tables

Pour créer les tables, nous avons trois choix : soit nous les créons en mode `création`, soit à l'aide de l'assistant ou encore en entrant les données. Nous choisissons ici le mode `création`.

Nous devons définir ici les noms des champs c'est-à-dire les noms des colonnes et le type de données qu'elles contiendront. Nous pouvons éventuellement entrer une description de ces colonnes.

Dans notre exemple, à partir de trois variables mesurées sur quinze individus, nous allons construire une base de données qui nous permettra au final d'obtenir une variable intervalle, une variable modale et une variable multivaluée mesurées sur cinq concepts (objets symboliques).

Nous définissons les champs suivants :

- IND : il s'agit de notre clé primaire qui nous permet d'identifier nos individus
- CONCEPTS : cette colonne nous permettra de construire nos objets symboliques
- INTER : il s'agit d'une variable quantitative qui a été mesurée sur les quinze individus
- MODAL : c'est une variable qualitative composée de trois catégories : A, B, C. Elle a également été mesurée sur les quinze individus.

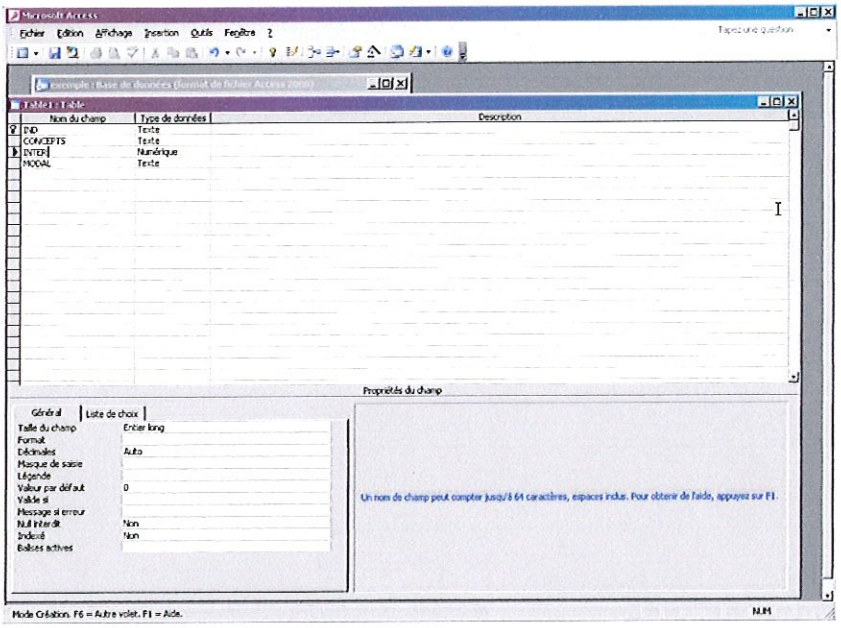


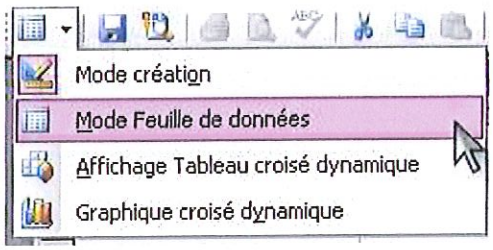
FIG. 7.27 – Construction d'une table en mode création

Nous devons ensuite enregistrer cette construction et donner un nom à notre table. Nous l'appelerons `TINDIVIDUS`.

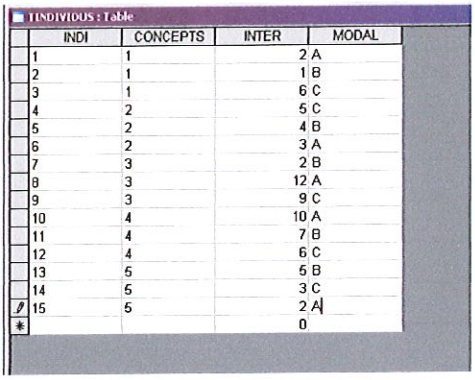
FIG. 7.28 – Enregistrement de la table

Pour entrer les valeurs dans notre tableau, nous pouvons soit les entrer directement soit utiliser les `formulaires`.

Nous choisissons ici de les entrer directement. Pour cela, nous passons en `mode feuille de données` par l'icône suivante :



Nous obtenons alors un tableau dans lequel nous pouvons entrer nos données.



INDI	CONCEPTS	INTER	MODAL
1	1	2	A
2	1	1	B
3	1	6	C
4	2	5	C
5	2	4	B
6	2	3	A
7	3	2	B
8	3	12	A
9	3	9	C
10	4	10	A
11	4	7	B
12	4	6	C
13	5	5	B
14	5	3	C
15	5	2	A
		0	

FIG. 7.29 – La table `TINDIVIDUS`

Pour créer des variables de type multi-catégoriques, nous devons définir un deuxième tableau `TCATEGORIES` à deux colonnes.

Ces colonnes sont :

- IND : il s'agit de la même colonne que dans notre tableau précédent
- CATEGORIES : il s'agit d'une variable qualitative à trois modalités : D, E et F.

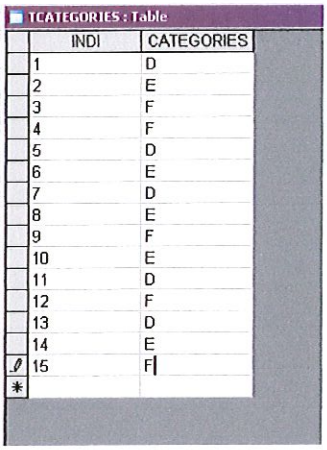


FIG. 7.30 – La table TCATEGORIES

Création des requêtes

Nous devons alors créer nos requêtes. Ces requêtes nous permettent de sélectionner certaines parties d'une table ou de réaliser une table à partir de plusieurs autres tables. Pour construire ces requêtes nous avons deux choix : soit le mode création soit la construction de requêtes à l'aide de l'assistant.

FIG. 7.31 – Création d'une requête

Comme précédemment, nous choisissons le mode création.

Remarques :

La requête que nous allons créer maintenant REQUETE_OS nous permettra, avec DB2SO, de créer nos variables symboliques intervalle et modale.

Nous devons tout d'abord choisir les tables qui apparaîtront dans la requête. Ici nous ne prendrons que `TINDIVIDUS`.

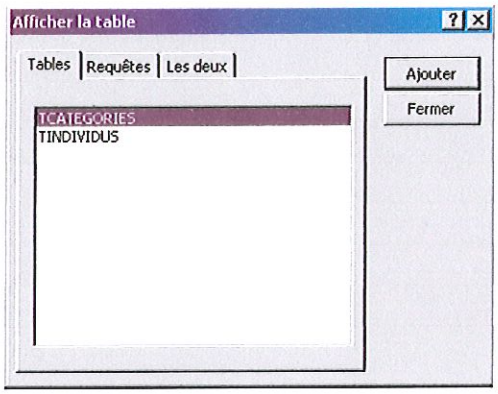


FIG. 7.32 – Choix des tables pour la requête

Nous devons ici choisir les identifiants, la variable qui nous permet de construire nos objets symboliques ainsi que les variables qui apparaîtront dans notre tableau de données symboliques. Il s'agit donc ici de toutes les colonnes de notre table `TINDIVIDUS`.

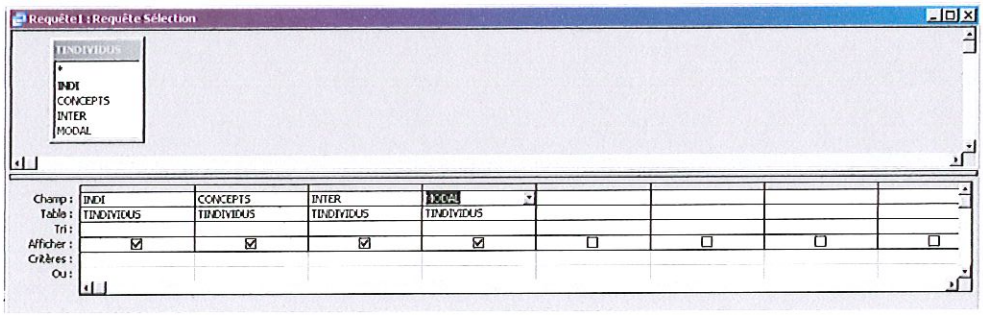


FIG. 7.33 – La requête `REQUETE_OS`

Remarque :

Les individus doivent absolument être dans la première colonne et que les concepts soient dans la deuxième.

Nous pouvons visualiser notre requête sous forme de table en sélectionnant le `mode feuille de données` comme nous l'avons fait précédemment.

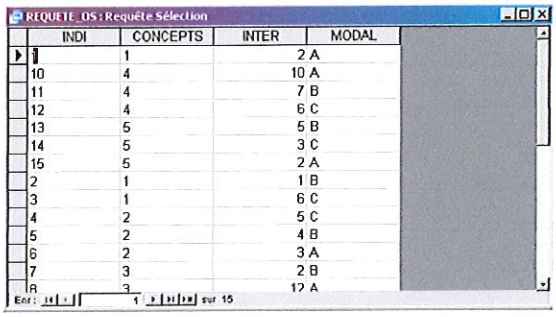


FIG. 7.34 – Visualisation de `REQUETE_OS`

Pour construire notre variable multivaluée, nous créons une deuxième requête `RMULTI`. Cette requête contiendra la table `TCATEGORIES`.

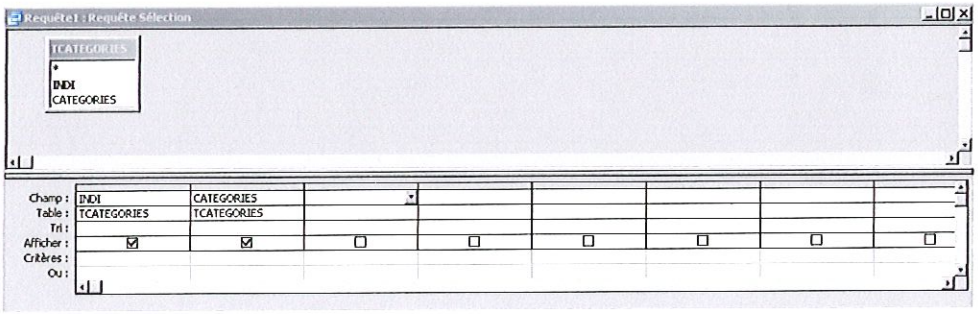


FIG. 7.35 – La requête `RMULTI`

La base de données `exemple.mdb`

Notre base de données finale contient donc 2 tables et 2 requêtes. Elle est finalement représentée de la manière suivante :

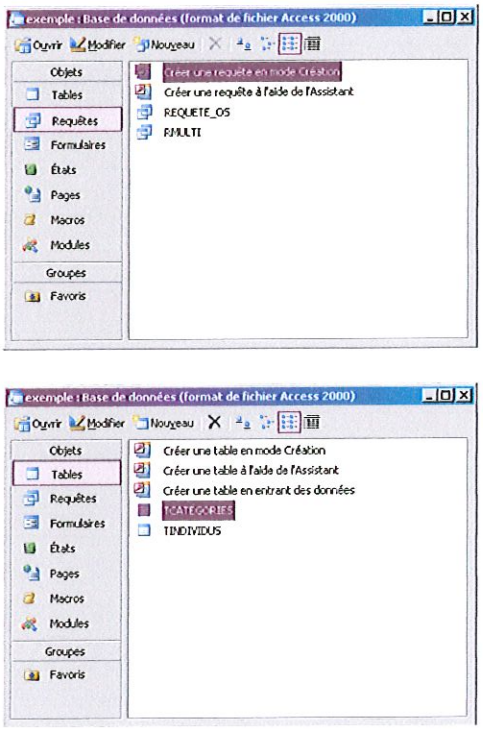


FIG. 7.36 – La base de données `exemple.mdb`

Annexe II : DB2SO

A partir d'une base de données Access, nous pouvons construire des objets symboliques à l'aide du logiciel DB2SO. Ce logiciel est inclus dans Sodas. Pour y accéder, il suffit d'aller dans l'onglet `Sodas file` et de sélectionner `Import...` suivi de `Import with DB2SO`.

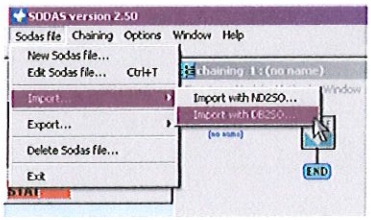


FIG. 7.37 – Chemin d'accès vers DB2SO

DB2SO se présente de la manière suivante :

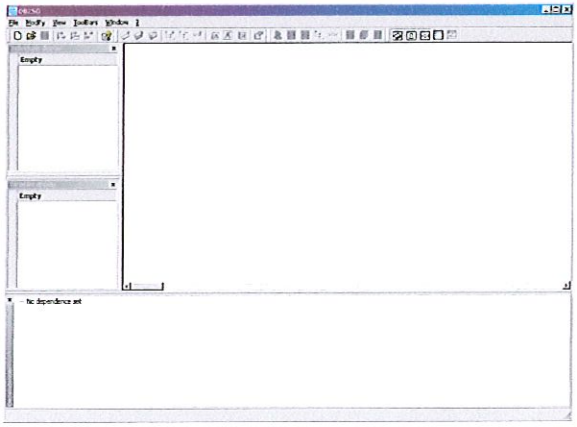


FIG. 7.38 – DB2SO

La première chose à faire est de sélectionner notre base de données. Pour cela, il faut sélectionner `New` dans l'onglet `open`. La fenêtre suivante s'ouvre alors :

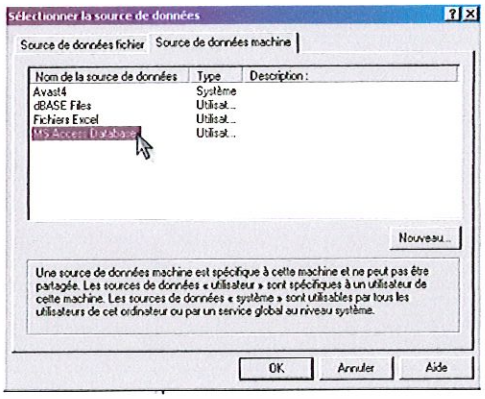


FIG. 7.39 – Choix du type de base de données

Puisque nous travaillons sur des bases de données Access, nous devons utiliser une source de données marchine `MS Access Database`. Une fois la source choisie, nous devons choisir la base de donnée Access sur laquelle nous souhaitons travailler. Nous prendrons ici la base de données `exemple.mdb`.

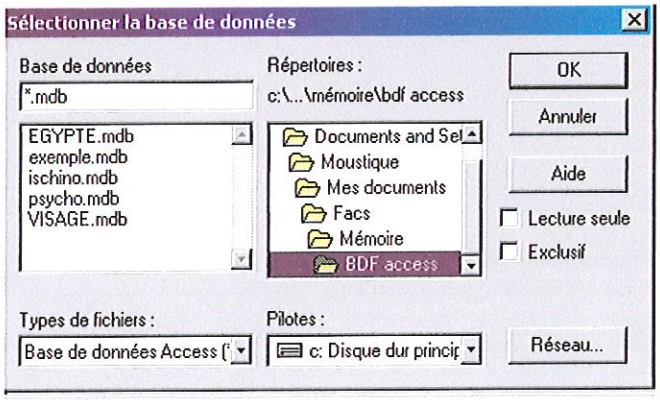


FIG. 7.40 – Choix de la base de données

Nous devons ensuite décider de la requête à exécuter. Cette requête correspond à celle qui a été définie dans la base de données Access pour construire nos objets symboliques. Il s'agit donc ici de la requête `REQUETE_OS`.

FIG. 7.41 – Choix de la requête à exécuter

L'exécution de cette requête par DB2SO nous permet d'obtenir une matrice de données symboliques.

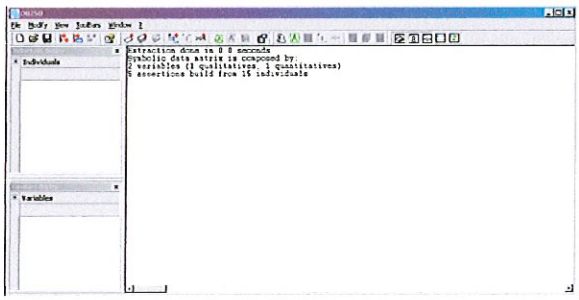


FIG. 7.42 – Construction de la matrice de données symboliques pour les données intervalles et modales

Remarquons cependant qu'à cette étape, seules les variables intervalles et modales sont construites.

Pour construire les données symboliques de type multivaluée, nous devons modifier la matrice symbolique obtenue à l'étape précédente. Pour cela il nous faut sélectionner `Add single-valued variables...` dans l'onglet `Modify`.

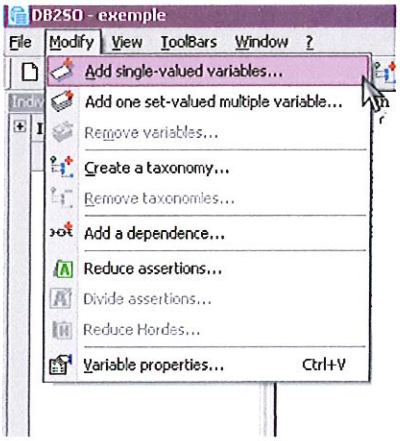


FIG. 7.43 – Modification de la matrice de données symbolique

Nous devons alors choisir la requête correspondant à notre variable catégorique : il s'agit ici de `RMULTI`

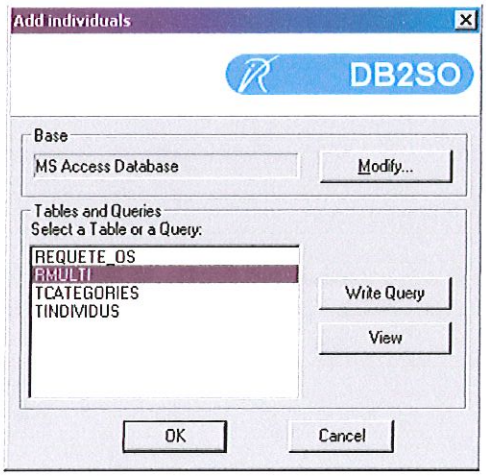


FIG. 7.44 – Choix de la requête pour la construction de la variable multivaluée

Nous obtenons alors une nouvelle matrice de données symboliques composée de trois variables : 2 variables qualitatives et une variable quantitative.

```
Extraction done in 0.0 seconds.
Symbolic data matrix is composed by:
2 variables (1 qualitatives, 1 quantitatives)
5 assertions build from 15 individuals
1 unique variables added:
  CATEGORIES
Total 3 variables (2 qualitatives, 1 quantitatives)
```

FIG. 7.45 – Matrice de données symboliques finale

Nous devons ensuite sauver cette matrice dans le format DB2SO c'est-à-dire avec l'extension `.gaj` (cette option se trouve dans l'onglet `File`).

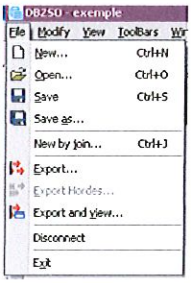


FIG. 7.46 – Onglet `File`

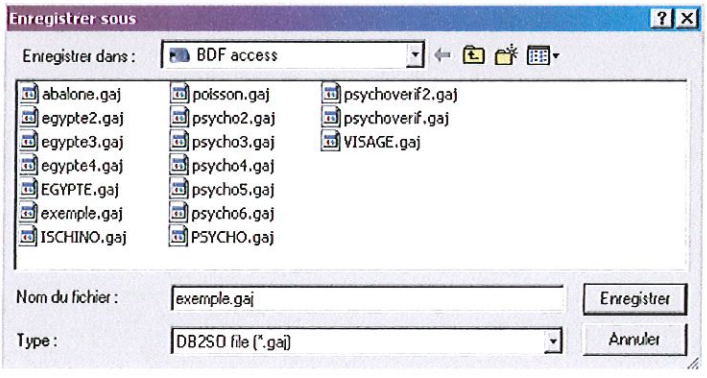


FIG. 7.47 – Enregistrement au format DB2SO

Pour obtenir notre base de données Sodas c'est-à-dire une base de données avec l'extension `.sds`, il ne nous reste plus qu'à exporter le fichier DB2SO (cette option se situe aussi dans l'onglet `File`).

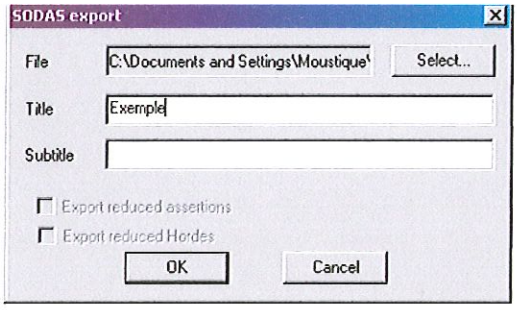


FIG. 7.48 – Exportation du fichier DB2SO

```
Extraction done in 0.0 seconds.
Symbolic data matrix is composed by:
2 variables (1 qualitatives, 1 quantitatives)
5 assertions build from 15 individuals
1 unique variables added:
  CATEGORIES
Total 3 variables (2 qualitatives, 1 quantitatives)
- Writing Meta Data file -
C:\Documents and Settings\Moustique\Mes documents\Facs\Mémoire\BDF access\exemple.gaj.xml
```

FIG. 7.49 – Sauvegarde et exportation de la matrice symbolique

Pour visualiser le tableau de données symboliques que nous avons construit, nous pouvons utiliser la méthode `View` proposée par Sodas.

	INTER	MODAL	CATEGORIES
AA00	[1.00 : 6.00]	A (0.33), B (0.33), C (0.33)	D
AA01	[3.00 : 5.00]	A (0.33), B (0.33), C (0.33)	E
AA02	[2.00 : 12.00]	A (0.33), B (0.33), C (0.33)	F
AA03	[6.00 : 10.00]	A (0.33), B (0.33), C (0.33)	F
AA04	[2.00 : 5.00]	A (0.33), B (0.33), C (0.33)	D

FIG. 7.50 – Tableau de données symboliques obtenu avec le fichier `exemple.sds`